SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 1–14, January 1993

001

IMPROVED ERROR BOUNDS FOR UNDERDETERMINED SYSTEM SOLVERS*

JAMES W. DEMMEL† AND NICHOLAS J. HIGHAM‡

Abstract. The minimal 2-norm solution to an underdetermined system $Ax = b$ of full rank can be computed using a QR factorization of A^T in two different ways. One method requires storage and reuse of the orthogonal matrix Q, while the method of seminormal equations does not. Existing error analyses show that both methods produce computed solutions whose normwise relative error is bounded to first order by $c\kappa_2(A)u$, where c is a constant depending on the dimensions of A, $\kappa_2(A) = \|A^+\|_2\|A\|_2$ is the 2-norm condition number, and u is the unit roundoff. It is shown that these error bounds can be strengthened by replacing $\kappa_2(A)$ by the potentially much smaller quantity $\mathrm{cond}_2(A) = \|\,|A^+|\cdot|A|\,\|_2$, which is invariant under row scaling of A. It is also shown that $\mathrm{cond}_2(A)$ reflects the sensitivity of the minimum norm solution x to row-wise relative perturbations in the data A and b. For square linear systems $Ax = b$ row equilibration is shown to endow solution methods based on LU or QR factorization of A with relative error bounds proportional to $\mathrm{cond}_\infty(A)$, just as when a QR factorization of A^T is used. The advantages of using fixed precision iterative refinement in this context instead of row equilibration are explained.

Key words. underdetermined system, seminormal equations, QR factorization, rounding error analysis, backward error, componentwise error bounds, iterative refinement, row scaling

AMS(MOS) subject classifications. 65F05, 65F25, 65G05

1. Introduction.

Consider the underdetermined system $Ax = b$, where $A \in \mathbb{R}^{m\times n}$ with $m \le n$. The system can be analysed using a QR factorization

$$A^T = Q\begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{1.1}$$

where $Q \in \mathbb{R}^{n\times n}$ is orthogonal and $R \in \mathbb{R}^{m\times m}$ is upper triangular. We have

$$b = Ax = [\,R^T \quad 0\,]\,Q^T x = R^T y_1, \tag{1.2}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = Q^T x.$$

If A has full rank then $y_1 = R^{-T}b$ is uniquely determined and all solutions of $Ax = b$ are given by

$$x = Q\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \qquad y_2 \in \mathbb{R}^{n-m} \text{ arbitrary}.$$

The unique solution x_{LS} that minimizes $\|x\|_2$ is obtained by setting $y_2 = 0$. We have

$$x_{LS} = Q\begin{bmatrix} R^{-T}b \\ 0 \end{bmatrix} \tag{1.3}$$

* Received by the editors September 4, 1990; accepted for publication July 18, 1991.

† Computer Science Division and Mathematics Department, University of California, Berkeley, California 94720 (na.demmel@na-net.ornl.gov). This author acknowledges the financial support of the National Science Foundation via grants DCR-8552474 and ASC-8715728. He is also a Presidential Young Investigator.

‡ Department of Mathematics, University of Manchester, Manchester, M13 9PL, United Kingdom (na.nhigham@na-net.ornl.gov).

$$\begin{aligned} &= Q\begin{bmatrix} R \\ 0 \end{bmatrix} R^{-1}R^{-T}b = Q\begin{bmatrix} R \\ 0 \end{bmatrix} (R^TR)^{-1}b \qquad (1.4)\\ &= A^T(AA^T)^{-1}b \\ &= A^+b, \end{aligned}$$

where $A^+ = A^T(AA^T)^{-1}$ is the pseudoinverse of A.

Equation (1.3) defines one way to compute x_{LS}. This is the method described in [13, Chap. 13], and we refer to it as the "Q method." When A is large and sparse it is desirable to avoid storing and accessing Q, which can be expensive. An alternative method with this property was suggested by Gill and Murray [6] and Saunders [16]. This method again uses the QR factorization (1.1) but computes x_{LS} as $x_{LS} = A^Ty$, where

$$R^TRy = b \qquad (1.5)$$

(cf. (1.4)). These latter equations are called the seminormal equations (SNE), since they are equivalent to the "normal equations" $AA^Ty = b$. As the "semi" denotes, however, this method does not explicitly form AA^T, which would be undesirable from the standpoint of numerical stability. We stress that equations (1.5) are different from the equations $R^TRx = A^Tb$ for an overdetermined least squares problem, where $A = Q[R^T \quad 0]^T \in \mathbb{R}^{m\times n}$ with $m \geq n$, yet these are also referred to as seminormal equations [4]. In this paper we are solely concerned with underdetermined systems so no confusion should arise.

Other methods for obtaining minimal 2-norm solutions of underdetermined systems are surveyed in [5].

Existing perturbation theory for the minimum norm solution problem, and error analysis for the above QR factorization-based methods, can be summarised as follows.

(1) Golub and Van Loan [7, Thm. 5.7.1] prove the following perturbation result. (Similar results are proved in [13, Thm. 9.18] and [20, Thm. 5.1].) Here, $\sigma_i(A)$ denotes the ith largest singular value of $A \in \mathbb{R}^{m\times n}$ and, if $\text{rank}(A) = m$, $\kappa_2(A) = \|A^+\|_2\|A\|_2 = \sigma_1(A)/\sigma_m(A)$.

THEOREM 1.1. *Let $A \in \mathbb{R}^{m\times n}$ and $0 \neq b \in \mathbb{R}^m$. Suppose that* $\text{rank}(A) = m \leq n$ *and that $\Delta A \in \mathbb{R}^{m\times n}$ and $\Delta b \in \mathbb{R}^m$ satisfy*

$$\epsilon = \max\{\|\Delta A\|_2/\|A\|_2, \|\Delta b\|_2/\|b\|_2\} < \sigma_m(A).$$

If x and $\widehat{x}$ are the minimum norm solutions to $Ax = b$ and $(A + \Delta A)\widehat{x} = b + \Delta b$, respectively, then

$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \leq \min\{3, n - m + 2\}\kappa_2(A)\epsilon + O(\epsilon^2). \qquad (1.6)$$

This result shows that small relative changes in the data A and b produce relative changes in the minimum norm solution x that are at most $\kappa_2(A)$ times as large. Unlike for the overdetermined least squares problem there is no term in $\kappa_2(A)^2$.

(2) Arioli and Laratta [2, Thm. 4] show that the computed solution $\widehat{x}$ from the Q method satisfies

$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \leq c_1u\kappa_2(A) + O(u^2), \qquad (1.7)$$

TABLE 1.1
Stability classification scheme.

	Backward stability	Forward stability
Normwise	N	N
Row-wise	R	R
Componentwise	C	C

where c_i denotes a modest constant depending on m and n, and u is the unit roundoff. (Arioli and Laratta actually analyse a slightly more general problem in which $\|x-w\|_2$ is minimized for a given vector w; we have taken $w=0$.)

(3) Paige [15] shows that the computed solution $\widehat{x}$ from the method of seminormal equations satisfies

$$\frac{\|\widehat{x}-x\|_2}{\|x\|_2} \le \kappa_2(A)c_1u + \frac{\kappa_2(A)c_2u(1+\kappa_2(A)c_3u)}{1-\kappa_2(A)c_4u}. \tag{1.8}$$

The bounds (1.7) and (1.8) are of the same form as (1.6). One implication of these existing results is that both the Q method and the SNE method are stable in the sense that the relative errors in the computed solutions reflect the sensitivity of the minimum norm problem to general perturbations in the data.

The purpose of this paper is to show that the results in (2) and (3) can be strengthened significantly by using componentwise analysis. First, in §2, we prove a version of Theorem 1.1 for componentwise perturbations; thus we measure ΔA and Δb by the smallest ϵ such that

$$|\Delta A| \le \epsilon E, \qquad |\Delta b| \le \epsilon f, \tag{1.9}$$

where $E \ge 0$ and $f \ge 0$ contain arbitrary tolerances and inequalities hold componentwise. We obtain an analogue of (1.6) with $\kappa_2(A)$ replaced by a potentially much smaller quantity that depends on A, x, E, and f.

In §3 we show that the term $\kappa_2(A)$ in (1.7) and (1.8) can be replaced by

$$\operatorname{cond}_2(A) \equiv \|\,|A^+|\cdot|A|\,\|_2,$$

which is a generalization of the condition number $\|\,|A^{-1}|\cdot|A|\,\|_2$ for square matrices introduced by Bauer [3] and Skeel [17]. This is important because $\operatorname{cond}_2(A)$ can be arbitrarily smaller than $\kappa_2(A)$, since $\operatorname{cond}_2(A)$ is invariant under row scalings $A \to DA$ (D diagonal and nonsingular), whereas $\kappa_2(A)$ is not. And $\operatorname{cond}_2(A)$ cannot be much bigger than $\kappa_2(A)$ since

$$\operatorname{cond}_2(A) \le \|\,|A^+|\,\|_2 \|\,|A|\,\|_2 \le n\|A^+\|_2\|A\|_2 = n\kappa_2(A). \tag{1.10}$$

In §§4 and 5 we investigate stability issues, and we encounter several different types of stability. To put these different types into perspective, we present a scheme that classifies six different kinds of stability in Table 1.1. (We appreciate that it can be counterproductive to over-formalize stability, but we believe that this scheme helps to clarify the overall picture.)

To explain the terminology we define for $A \in \mathbb{R}^{m\times n}$, with $m \le n$, the backward error

$$\begin{aligned}\omega_{E,f}(y) \equiv \min\{\epsilon :\ & \exists\ \Delta A \in \mathbb{R}^{m\times n}, \Delta b \in \mathbb{R}^m \text{ s.t. } y \text{ is the minimum norm} \\ & \text{solution to } (A+\Delta A)y = b+\Delta b, \text{ and } |\Delta A| \le \epsilon E,\ |\Delta b| \le \epsilon f\},\end{aligned}$$

where $E \geq 0$ and $f \geq 0$ are given. Note that if we were to remove the minimum norm requirement on y in the definition of $\omega_{E,f}$ then the backward error would be given by

$$\max_i \frac{|b - Ay|_i}{(E|y| + f)_i}, \tag{1.11}$$

as shown in [14]. The three measures of backward stability in Table 1.1 correspond to the following choices of E and f, where $e_n = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$:

$$\begin{aligned} \text{normwise } (\omega^N): &\quad E_N = \|A\|_2 e_m e_n^T, \qquad f_N = \|b\|_2 e_m, \\ \text{row-wise } (\omega^R): &\quad E_R = |A| e_n e_n^T, \quad f_R = |b|, \\ \text{componentwise } (\omega^C): &\quad E_C = |A|, \quad f_C = |b|. \end{aligned} \tag{1.12}$$

A small value for $\omega^R(y)$ means that y is the minimum norm solution to a perturbed system where the perturbation to the ith row of A is small compared with the norm of the ith row (similarly for b). We say, for example, that a numerical method for solving $Ax = b$ is in backward stability category R (or is row-wise backward stable) if it produces a computed solution $\widehat{y}$ such that $\omega^R(\widehat{y})$ is of order the unit roundoff.

For each type of backward error there is a perturbation result that bounds $\|x - y\|_2/\|x\|_2$ by a multiple of $\omega_{E,f}(y)$, and the multiplier defines a condition number. As explained in §2, for underdetermined systems the conditions numbers are $\kappa_2(A)$ for ω^N, $\text{cond}_2(A)$ for ω^R, and a quantity $\text{cond}_2(A, x)$ that depends on both A and x for ω^C. Continuing the "R-stability" example above, we say that a method is in forward stability category R if it has a forward error bound of order $\text{cond}_2(A)$ times the unit roundoff. An algorithm that has backward stability X (where $X = N$, R, or C) automatically has forward stability X; one of the reasons these definitions are useful is that an algorithm can have forward stability X without having backward stability X.

In this terminology, the gist of §3 is that the Q method and the SNE method have forward stability R, whereas previous results guaranteed only forward stability N.

In §4 we explain why the Q method is (nearly) row-wise backward stable but the SNE method is not backward stable at all. We give some numerical results to provide insight into the error bounds, and to illustrate the performance of fixed precision iterative refinement with the SNE method.

In §5 we consider the implications of the results of §3 for square linear systems. We show that row equilibration of the system $Ax = b$ allows methods based on LU and QR factorization of A to produce computed solutions whose relative errors are bounded in the same way as when a QR factorization of A^T is used—namely by a multiple of $\text{cond}(A)u$ (corresponding to row-wise forward stability). We explain why fixed precision iterative refinement leads to an even more satisfactory computed solution than row equilibration, and we provide two numerical examples for illustration.

2. Componentwise perturbation result. In this section we prove the following componentwise perturbation result for the minimum norm problem, and use it to determine the condition numbers for the perturbation measures in (1.12).

Theorem 2.1. *Let* $A \in \mathbb{R}^{m \times n}$ *and* $0 \neq b \in \mathbb{R}^m$. *Suppose that* $\text{rank}(A) = m \leq n$, *and that*

$$|\Delta A| \leq \epsilon E, \qquad |\Delta b| \leq \epsilon f,$$

where $E \geq 0$, $f \geq 0$, and $\epsilon\|E\|_2 < \sigma_m(A)$. If x and $\widehat{x}$ are the minimum norm solutions to $Ax = b$ and $(A + \Delta A)\widehat{x} = b + \Delta b$, respectively, then

(2.1)
$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \leq (\| \, |I - A^+A| \cdot E^T \cdot |A^{+^T} x| \, \|_2 + \| \, |A^+| \cdot (f + E|x|) \, \|_2) \frac{\epsilon}{\|x\|_2} + O(\epsilon^2).$$

Proof. $A + \Delta A$ has full rank so we can manipulate the equation

$$\widehat{x} = (A + \Delta A)^T ((A + \Delta A)(A + \Delta A)^T)^{-1} (b + \Delta b)$$

to obtain

$$\begin{aligned} \widehat{x} - x &= (I - A^+A)\Delta A^T (AA^T)^{-1} b + A^+(\Delta b - \Delta A x) + O(\epsilon^2) \\ &= (I - A^+A)\Delta A^T A^{+^T} x + A^+(\Delta b - \Delta A x) + O(\epsilon^2). \end{aligned} \tag{2.2}$$

Taking norms and then using absolute value inequalities, together with the monotonicity property $|x| \leq y \Rightarrow \|x\|_2 \leq \|y\|_2$, we have

$$\begin{aligned} \|\widehat{x} - x\|_2 &\leq \|(I - A^+A)\Delta A^T A^{+^T} x\|_2 + \|A^+(\Delta b - \Delta A x)\|_2 + O(\epsilon^2) \\ &\leq (\| \, |I - A^+A| \cdot E^T \cdot |A^{+^T} x| \, \|_2 + \| \, |A^+| \cdot (f + E|x|) \, \|_2)\epsilon + O(\epsilon^2), \end{aligned}$$

as required. □

We note that for given A, b, E, and f there exist ΔA and Δb for which the bound in (2.1) is attained to within a constant factor depending on n. This is a consequence of the fact that the two vectors on the right-hand side of (2.2) are orthogonal. Also, it is clear from the proof that (2.1) is valid with the 2-norm replaced by the ∞-norm.

By substituting the E and f from (1.12) into Theorem 2.1 we can deduce the condition numbers corresponding to our three different ways of measuring the perturbations ΔA and Δb. For the componentwise measure the condition number is clearly

$$\begin{aligned} \text{cond}_2(A, x) = (\| \, &|I - A^+A| \cdot |A^T| \cdot |A^{+^T} x| \, \|_2 \\ &+ \| \, |A^+| \cdot (|b| + |A||x|) \, \|_2) / \|x\|_2. \end{aligned} \tag{2.3}$$

Replacing b by its upper bound $|A||x|$ simplifies this expression while increasing it by no more than a factor of 2.

For the row-wise measure the bracketed term in the bound in (2.1) is within a factor depending on n of $\text{cond}_2(A)$, hence we can take $\text{cond}_2(A)$ as the condition number. In showing this, we need to use the equality $\|I - A^+A\|_2 = \min\{1, n - m\}$ (which can be derived by consideration of the QR factorization (1.1), for example), and the observation that if $B \in \mathbb{R}^{m \times n}$ and $B \geq 0$, then

$$\frac{1}{\sqrt{m}} \|B\|_2 \leq \|B\|_\infty = \|Be\|_\infty \leq \|Be\|_2 \leq \sqrt{n} \|B\|_2.$$

Note that when $|x| = e$, $\text{cond}_2(A)$ differs from $\text{cond}_2(A, x)$ by no more than a factor of about $\sqrt{n}$. Finally, for the normwise measure the condition number is $\kappa_2(A)$ (as implied by Theorem 1.1). Table 2.1 summarises these results.

TABLE 2.1
Condition numbers.

Measure	Condition number
Normwise	$\kappa_2(A)$
Row-wise	$\mathrm{cond}_2(A)$
Componentwise	$\mathrm{cond}_2(A,x)$

In the error analysis of the next section we need to use Theorem 2.1 with $E = |A|H$, where H is a given matrix. In this case, taking also $f = |b|$, it is convenient to put (2.1) in the form

$$\frac{\|\widehat{x}-x\|_2}{\|x\|_2} \le \min\{3, n-m+2\}\max\{\|H\|_2, 1\}\mathrm{cond}_2(A)\epsilon + O(\epsilon^2). \tag{2.4}$$

If $\|H\|_2 = 1$, this is precisely (1.6) with $\kappa_2(A)$ replaced by $\mathrm{cond}_2(A)$, this difference reflecting the stronger assumption made about the perturbations for (2.4).

3. Error analysis. In this section we carry out an error analysis of the Q method and the SNE method. We assume that the floating point arithmetic obeys the model

$$\begin{aligned} fl(x \,\mathrm{op}\, y) &= (x \,\mathrm{op}\, y)(1+\delta), \qquad |\delta| \le u, \quad \mathrm{op} = *, /,\\ fl(x \pm y) &= x(1+\alpha) \pm y(1+\beta), \qquad |\alpha|, |\beta| \le u,\\ fl(\sqrt{x}) &= \sqrt{x}(1+\delta), \qquad |\delta| \le u. \end{aligned}$$

We consider first the Q method, and we assume that the QR factorization (1.1) is computed by Householder transformations or Givens transformations. In [12, Cor. A.8] it is shown that if $\widehat{R}$ is the computed upper triangular factor, there exists an orthogonal matrix $\widetilde{Q}$ such that

$$A^T + \Delta A^T = \widetilde{Q}\begin{bmatrix} \widehat{R} \\ 0 \end{bmatrix}, \tag{3.1}$$

where

$$|\Delta A^T| \le G_{m,n} u |A^T| \tag{3.2}$$

and $\|G_{m,n}\|_2 \le \mu_{m,n}$. Here and below we use $\mu_{m,n}$ generically to denote a modest constant depending on m and n; we are not concerned with the precise values of the constants so we will freely write, for example, $\mu_{m,n} + \mu'_{m,n} = \mu''_{m,n}$.

The Q method solves the triangular system $R^T y_1 = b$ and forms $x = Q\,[\,y_1^T,\ 0\,]^T$. Standard analysis shows that the computed $\widehat{y}_1$ satisfies

$$(\widehat{R} + \Delta\widehat{R})^T \widehat{y}_1 = b, \qquad |\Delta\widehat{R}| \le \mu_m u |\widehat{R}|. \tag{3.3}$$

From [12, Lemma A.7] the computed solution $\widehat{x}$ satisfies

$$\widehat{x} = \widetilde{Q}\begin{bmatrix} \widehat{y}_1 \\ 0 \end{bmatrix} + g, \tag{3.4}$$

where

$$|g| \le G'_{m,n}\begin{bmatrix} |\widehat{y}_1| \\ 0 \end{bmatrix} u, \qquad \|G'_{m,n}\|_2 \le \mu'_{m,n}. \tag{3.5}$$

(We emphasise the important point that the same orthogonal matrix $\widetilde{Q}$ appears in (3.1) and (3.4).)

Ideally, we would like to use the basic error equations (3.1)–(3.5) to show that $\widehat{x}$ is the exact minimum norm solution to a perturbed problem where the perturbations are bounded according to $|\Delta A| \le \epsilon |A|$ and $|\Delta b| \le \epsilon |b|$. The forward error could then be bounded by invoking (2.1). Unfortunately, this componentwise backward stability result does not hold. We can, nevertheless, obtain a forward error bound of the form (2.4) by using a mixed forward and backward error argument.

From (3.3), (3.4), and (3.1) we have

$$\begin{aligned} b &= \begin{bmatrix} (\widehat{R} + \Delta\widehat{R})^T & 0 \end{bmatrix} \widetilde{Q}^T \cdot \widetilde{Q} \begin{bmatrix} \widehat{y}_1 \\ 0 \end{bmatrix} \\ &= (A + F)\overline{x}, \end{aligned}$$

where

$$\begin{aligned} F &= \Delta A + \begin{bmatrix} \Delta\widehat{R}^T & 0 \end{bmatrix} \widetilde{Q}^T, \\ \overline{x} &= \widetilde{Q} \begin{bmatrix} \widehat{y}_1 \\ 0 \end{bmatrix}. \end{aligned} \tag{3.6}$$

Since $(A+F)^T$ has the QR factorization $(A+F)^T = \widetilde{Q} \begin{bmatrix} (\widehat{R} + \Delta\widehat{R})^T & 0 \end{bmatrix}^T$ it follows from (3.3) and (3.6) that $\overline{x}$ is the minimum norm solution to $(A+F)\overline{x} = b$ as long as $\|F\|_2 < \sigma_m(A)$ (so that $A+F$ has full rank). From (3.1)–(3.3) we have

$$\begin{aligned} |F| &\le u|A|G_{m,n}^T + \mu_m u|A|(I + uG_{m,n}^T)|\widetilde{Q}||\widetilde{Q}^T| \\ &\equiv u|A|H_{m,n}. \end{aligned}$$

Hence we can invoke (2.4) to obtain

$$\frac{\|\overline{x} - x\|_2}{\|x\|_2} \le \mu_{m,n} \mathrm{cond}_2(A) u + O(u^2). \tag{3.7}$$

Now from (3.4), (3.5), and (3.6) we have

$$\begin{aligned} \|\widehat{x} - \overline{x}\|_2 &= \|g\|_2 \\ &\le \mu'_{m,n} \|\widehat{y}_1\|_2 u = \mu'_{m,n} \|\widehat{x}\|_2 u + O(u^2) \\ &= \mu'_{m,n} \|x\|_2 u + O(u^2). \end{aligned} \tag{3.8}$$

Combining (3.7) and (3.8) we conclude that

$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \le \mu''_{m,n} \mathrm{cond}_2(A) u + O(u^2). \tag{3.9}$$

Now we analyse the SNE method. As for the Q method, (3.1) and (3.2) hold for the computed triangular factor $\widehat{R}$. The computed solution $\widehat{y}$ to (1.5) satisfies

$$(\widehat{R} + \Delta\widehat{R}_1)^T (\widehat{R} + \Delta\widehat{R}_2)\widehat{y} = b, \qquad |\widehat{R}_i| \le \mu_m u |\Delta\widehat{R}_i|, \tag{3.10}$$

and the computed solution $\widehat{x}$ satisfies

$$\widehat{x} = A^T\widehat{y} + g, \qquad |g| \le \mu_m u |A^T||\widehat{y}|. \tag{3.11}$$

Taking a similar approach to the analysis for the Q method we write

$$\widehat{x} = \overline{x} + \Delta\overline{x}, \tag{3.12}$$

where

$$\overline{x} = (A + \Delta A)^T\overline{y},$$

$$\widehat{R}^T\widehat{R}\overline{y} = b, \tag{3.13}$$

$$\Delta\overline{x} = A^T(\widehat{y} - \overline{y}) - \Delta A^T\overline{y} + g. \tag{3.14}$$

Note that $\overline{x}$ is the exact minimum norm solution to $(A + \Delta A)x = b$ and so once again (3.7) holds.

For later use we note that from (3.1),

$$A + \Delta A = \widehat{R}^T\widetilde{Q}_1^T, \tag{3.15}$$

where $\widetilde{Q}_1$ comprises the first m columns of $\widetilde{Q}$, and hence, using (3.13),

$$\widetilde{Q}_1^T\overline{x} = \widehat{R}\overline{y}. \tag{3.16}$$

It remains to bound $\Delta\overline{x}$. Straightforward manipulation of (3.10) and (3.13) yields

$$\begin{aligned} \overline{y} - \widehat{y} &= \widehat{R}^{-1}\widehat{R}^{-T}\Delta\widehat{R}_1^T\widehat{R}\overline{y} + \widehat{R}^{-1}\Delta\widehat{R}_2\overline{y} + O(u^2) \\ &= \widehat{R}^{-1}(\widehat{R}^{-T}\Delta\widehat{R}_1^T\widetilde{Q}_1^T\overline{x} + \Delta\widehat{R}_2\overline{y}) + O(u^2), \end{aligned}$$

where we have used (3.16). Premultiplying by A^T and using (3.15) gives

$$A^T(\overline{y} - \widehat{y}) = \widetilde{Q}_1\big(\widehat{R}^{-T}\Delta\widehat{R}_1^T\widetilde{Q}_1^T\overline{x} + \Delta\widehat{R}_2\overline{y}\big) + O(u^2),$$

which leads to

$$\|A^T(\overline{y} - \widehat{y})\|_2 \le \mu_m u\big(\|\,|\widehat{R}^{-T}|\cdot|\widehat{R}^T|\,\|_2\|\overline{x}\|_2 + \|\,|\widehat{R}|\cdot|\overline{y}|\,\|_2\big) + O(u^2). \tag{3.17}$$

To bound $\|\,|\widehat{R}^{-T}|\cdot|\widehat{R}^T|\,\|_2$, note that for the exact QR factorization we have

$$\|\,|R^{-T}|\cdot|R^T|\,\|_2 = \|\,|Q_1^TA^+|\cdot|AQ_1|\,\|_2 \le m\,\mathrm{cond}_2(A).$$

Hence

$$\|\,|\widehat{R}^{-T}|\cdot|\widehat{R}^T|\,\|_2 \le m\,\mathrm{cond}_2(A + \Delta A) = m\,\mathrm{cond}_2(A) + O(u). \tag{3.18}$$

To bound $\|\,|\widehat{R}|\cdot|\overline{y}|\,\|_2$ in (3.17) we note first that for the exact R and y,

$$\|\,|R|\cdot|y|\,\|_2 = \|\,|Q_1^TA^T|\cdot|y|\,\|_2 \le \sqrt{m}\|\,|A^T|\cdot|y|\,\|_2.$$

Now, since $x = A^Ty$, we have $Ax = (AA^T)y$ or $y = A^{+^T}x$. Hence

$$\|\,|A^T|\cdot|y|\,\|_2 \le \|\,|A^T|\cdot|A^{+^T}|\cdot|x|\,\|_2 \le \mathrm{cond}_2(A)\|x\|_2. \tag{3.19}$$

It follows that for the computed $\widehat{R}$ and $\overline{y}$,

$$\|\,|\widehat{R}|\cdot|\overline{y}|\,\|_2 \le \sqrt{m}\,\mathrm{cond}_2(A)\|x\|_2 + O(u). \tag{3.20}$$

Combining (3.14), (3.17), (3.18), (3.20), (3.11), and (3.19) we have

$$\|\Delta\overline{x}\|_2 \le \mu_{m,n}\mathrm{cond}_2(A)u\|x\|_2 + O(u^2).$$

Together with (3.7) and (3.12) this yields

$$\frac{\|\widehat{x} - x\|_2}{\|x\|_2} \le \mu'_{m,n}\mathrm{cond}_2(A)u + O(u^2).$$

4. Discussion and numerical results. The analysis in the previous section shows that for both the Q method and the SNE method the forward error is bounded by a multiple of $\mathrm{cond}_2(A)u$, so both methods are forward stable in the row-wise sense. Before giving some numerical examples we briefly consider what can be said about backward stability.

For the Q method, the analysis of §3 proves the following result about the computed solution $\widehat{x}$. There exists a vector $\overline{x}$ and a matrix F such that $\overline{x}$ is the minimum norm solution to $(A+F)\overline{x} = b$ where

$$|F| \le u|A|H_{m,n} \le \mu_{m,n}u|A|ee^T \quad \Rightarrow \quad \|F\|_2 \le \mu'_{m,n}u\|A\|_2$$

and

$$\|\widehat{x} - \overline{x}\|_2 \le \mu''_{m,n}\|\widehat{x}\|_2 u + O(u^2).$$

(This result, without the componentwise bound on F, is also proved in [13, Thm. 16.18].) Thus $\widehat{x}$ is relatively close to a vector that satisfies the criterion for row-wise backward stability, and so the Q method is "almost" row-wise backward stable. Note also that, from the above, $\widehat{x}$ has a relatively small residual:

$$\|b - A\widehat{x}\|_2 \le \mu'''_{m,n}\|A\|_2\|\widehat{x}\|_2 u + O(u^2). \tag{4.1}$$

Interestingly, (4.1) implies that $\widehat{x}$ itself solves a slightly perturbed system, but it is not, in general, the minimum norm solution.

For the SNE method it is not even possible to derive a residual bound of the form (4.1). The method of solution guarantees only that the seminormal equations themselves have a small residual. Thus, as in the context of overdetermined least squares problems [4], the SNE method is not backward stable.

A possible way to improve the backward stability of the SNE method is to use iterative refinement in fixed precision, as advocated in the overdetermined case in [4]. Some justification for this approach can be given using the analysis for an arbitrary linear equations solver in [12].

We have run some numerical experiments in MATLAB, which has a unit roundoff $u \approx 2.2 \times 10^{-16}$. In our experiments we rounded the result of every arithmetic operation to 23 significant bits, thus simulating single precision arithmetic with $u_{\mathrm{SP}} \approx 1.2 \times 10^{-7}$. The double precision solution was regarded as the exact solution when computing forward errors.

We report results for several 10×16 matrices A, with the right-hand sides b chosen randomly with elements from the normal $(0,1)$ distribution. We report for each approximate solution $\widehat{y}$, the normwise relative error

$$\gamma_2(\widehat{y}) = \frac{\|\widehat{y} - x\|_2}{\|x\|_2},$$

and the three relative residuals

$$\rho^X(\widehat{y}) = \max_i \frac{|b - A\widehat{y}|_i}{(E_X|\widehat{y}| + f_X)_i}, \qquad X = N, R, C,$$

where E_X and f_X are defined in (1.12). Iterative refinement in fixed precision was used with the SNE method until either $\rho^N(\widehat{y}) \le u_{\mathrm{SP}}$ or five iterations were done. Note that if we were to use the ∞-norm in defining E_N and f_N in (1.12), then

$\rho^N(\widehat{y}) \le \rho^R(\widehat{y})$ would be guaranteed; for the 2-norm, $\rho^N(\widehat{y}) > \rho^R(\widehat{y})$ is possible. We also report the three condition numbers for each problem. There is no strict ordering between these condition numbers (partly, again, because of the choice of norm), but there are constants c_1 and c_2 depending only on n such that

$$\mathrm{cond}_2(A,x) \le c_1 \mathrm{cond}_2(A) \le c_2 \kappa_2(A)$$

(see (1.10) and §2).

The results are presented in Tables 4.1–4.6. The matrices A in Tables 4.1–4.3 are random matrices with geometrically distributed singular values $\sigma_i = \alpha^i$, generated using the routine randsvd of [10]. In Table 4.4, $Ax = b$ is the same system used in Table 4.1, but with the fifth equation scaled by $2^{15} = 32768$. In Table 4.5 the system is the one used in Table 4.1 but with the eighth column of A scaled by 2^{15}. In Table 4.6, A is a Kahan matrix—an ill-conditioned upper trapezoidal matrix with rows of widely varying norm [7, p. 245], [10].

The key features in the results are as follows.

(1) The error bounds of the previous section are confirmed. Indeed, for both the Q method and the SNE method the heuristic

$$\gamma_2(\widehat{x}) = \frac{\|\widehat{x} - x\|_2}{\|x\|_2} \approx \mathrm{cond}_2(A)u$$

predicts the error correctly to within an order of magnitude in these examples.

(2) The independence of the forward errors on the row scaling of A is illustrated by Tables 4.1 and 4.4. However, column scaling can have an adverse effect, as shown in Table 4.5.

(3) The relative residuals confirm that the Q method is (almost) row-wise backward stable and that the SNE method is not even normwise backward stable. The relative residuals for the SNE method exhibit dependence on $\mathrm{cond}_2(A)$ in these examples (dependence of the normwise residual on $\kappa_2(A)$ in the case of overdetermined systems is proven by Björck in [4, Thm. 3.1]). Iterative refinement can produce a small relative residual, but can fail on very ill conditioned problems, as in Table 4.3.

The condition numbers displayed in the tables can all be estimated cheaply given a QR factorization of A^T. For example, we show how to estimate $\mathrm{cond}_2(A,x)$. This differs by at most a factor $\sqrt{n}$ from $\mathrm{cond}_\infty(A,x)$. We consider only the first term of $\mathrm{cond}_\infty(A,x)$ in (2.3), as the second term can be treated similarly. As in [1], we can convert this norm of a vector into a norm of a matrix: with $g = |A^T||A^{+^T}x|$ and $G = \mathrm{diag}(g_i)$, we have

$$\begin{aligned}\|\,|I - A^+A| \cdot |A^T| \cdot |A^{+^T}x|\,\|_\infty &= \|\,|I - A^+A|g\,\|_\infty \\ &= \|\,|I - A^+A|Ge\,\|_\infty = \|\,|I - A^+A|G\,\|_\infty \\ &= \|\,|(I - A^+A)G|\,\|_\infty \\ &= \|\,(I - A^+A)G\,\|_\infty.\end{aligned}$$

The latter norm can be estimated by the method of [8], [9], and [11], which estimates $\|B\|_1$ given a means for forming matrix–vector products Bx and B^Ty. Forming these products for $B^T = (I - A^+A)G$ involves multiplying by G and Q or their transposes, and solving triangular systems with R and R^T.

TABLE 4.1
$A = \text{randsvd}([10,16], 1\text{e}2)$, $\kappa_2(A) = 1\text{e}2$, $\text{cond}_2(A) = 8.63\text{e}1$, $\text{cond}_2(A,x) = 1.57\text{e}2$.

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	1.83e−8	9.88e−9	1.42e−7	2.01e−6
SNE	5.11e−7	2.79e−7	4.40e−6	4.97e−6
	1.52e−8	6.45e−9	9.64e−8	1.99e−6

TABLE 4.2
$A = \text{randsvd}([10,16], 1\text{e}4)$, $\kappa_2(A) = 1\text{e}4$, $\text{cond}_2(A) = 5.43\text{e}3$, $\text{cond}_2(A,x) = 1.05\text{e}4$.

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	5.16e−9	6.84e−9	9.16e−8	1.29e−4
SNE	1.30e−5	1.56e−5	2.36e−4	2.30e−4
	4.29e−9	4.63e−9	8.01e−8	1.04e−4

TABLE 4.3
$A = \text{randsvd}([10,16], 1\text{e}6)$, $\kappa_2(A) = 1\text{e}6$, $\text{cond}_2(A) = 4.30\text{e}5$, $\text{cond}_2(A,x) = 8.86\text{e}5$.

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	6.88e−9	5.78e−9	9.18e−8	6.50e−3
SNE	3.58e−3	1.69e−3	2.56e−2	2.47e−2
	5.17e−5	2.47e−5	3.74e−4	1.28e−2
	5.39e−6	2.62e−6	3.96e−5	1.11e−2
	2.05e−5	9.33e−6	1.41e−4	1.11e−2
	1.51e−5	6.94e−6	1.05e−4	1.27e−2

TABLE 4.4
$A = D\,\text{randsvd}([10,16], 1\text{e}2)$, $\kappa_2(A) = 1.63\text{e}6$, $\text{cond}_2(A) = 8.63\text{e}1$, $\text{cond}_2(A,x) = 1.57\text{e}2$.

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	9.24e−9	9.88e−9	1.42e−7	2.01e−6
SNE	9.26e−7	2.79e−7	4.40e−6	4.97e−6
	5.70e−9	6.45e−9	9.64e−8	1.99e−6

5. Implications for square linear systems.

All the results in §§2 and 3 are valid when $m = n$. Theorem 2.1 reduces to a straightforward generalization of a result in [17, Thm. 2.1]. However, the error bound

$$\frac{\|\widehat{x} - x\|_\infty}{\|x\|_\infty} \le \mu_n \text{cond}_\infty(A) u + O(u^2) \tag{5.1}$$

for the Q method is not a familiar one for square systems. (We have switched to the ∞-norm, which is the more usual choice for square systems.) In fact, a bound of the form (5.1) holds also if we solve $Ax = b$ using an LU factorization (with partial pivoting) of A^T. Of course, when solving a square system $Ax = b$, it is more natural to use an LU or QR factorization of A than of A^T. But if a factorization of A is used, then no bound of the form (5.1) holds in general—the best we can say is that

$$\frac{\|\widehat{x} - x\|_\infty}{\|x\|_\infty} \le \mu_n \kappa_\infty(A) u + O(u^2). \tag{5.2}$$

We note, however, that there is a simple way to achieve a bound of the form (5.1) for LU and QR factorization of A: work with the scaled system $(DA)x = Db$ instead of $Ax = b$, where $B = DA$ has rows of unit 1-norm. This follows from (5.2)

TABLE 4.5
$A = \text{randsvd}([10,16], 1\text{e}2)\, D$, $\kappa_2(A) = 1.37\text{e}6$, $\text{cond}_2(A) = 7.81\text{e}5$, $\text{cond}_2(A,x) = 1.35\text{e}2$.

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	2.42e−9	5.70e−9	2.95e−4	9.29e−3
SNE	4.23e−3	5.89e−3	9.98e−1	2.61e−2
	1.39e−5	1.93e−5	5.89e−1	1.75e−3
	4.24e−7	5.90e−7	4.34e−2	3.60e−5
	3.02e−9	4.20e−9	3.12e−4	1.29e−6

TABLE 4.6
$A = \text{kahan}([10,16])$, $\kappa_2(A) = 6.29\text{e}5$, $\text{cond}_2(A) = 9.58\text{e}0$, $\text{cond}_2(A,x) = 1.02\text{e}1$.

	$\rho^N(\widehat{y})$	$\rho^R(\widehat{y})$	$\rho^C(\widehat{y})$	$\gamma_2(\widehat{y})$
Q method	1.22e−8	3.52e−9	4.99e−8	1.79e−7
SNE	8.00e−8	3.42e−8	3.27e−7	3.35e−7

and the fact that $\kappa_\infty(B) = \text{cond}_\infty(A)$. To verify the latter equality note that if $D^{-1} = \text{diag}(|A|e)$, then

$$\begin{aligned}\text{cond}_\infty(A) &= \|\, |A^{-1}| \cdot |A| \,\|_\infty = \|\, |A^{-1}| \cdot |A|e \,\|_\infty = \|\, |A^{-1}|D^{-1}e \,\|_\infty \\ &= \|\, |A^{-1}|D^{-1} \,\|_\infty = \|\, |A^{-1}D^{-1}| \,\|_\infty = \|\, |(DA)^{-1}| \,\|_\infty \\ &= \|\, B^{-1} \,\|_\infty = \kappa_\infty(B).\end{aligned}$$

It is interesting to compare this row equilibration strategy with fixed precision iterative refinement (FPIR). It is known that under suitable assumptions, FPIR in conjunction with LU factorization with partial pivoting [1], [18] or QR factorization [12] leads to a computed $\widehat{y}$ such that $\omega^C(\widehat{y}) = O(u)$, that is, FPIR brings componentwise backward stability. From an ∞-norm version of Theorem 2.1 we see that $\omega^C(\widehat{y}) \le u$ implies

$$\frac{\|\widehat{y} - x\|_\infty}{\|x\|_\infty} \le 2\,\text{cond}_\infty(A,x)u + O(u^2),$$

where

$$\text{cond}_\infty(A,x) = \frac{\|\, |A^{-1}| \cdot |A| \cdot |x| \,\|_\infty}{\|x\|_\infty}.$$

This is a stronger bound than (5.1) because $\text{cond}_\infty(A,x) \le \text{cond}_\infty(A)$ (with equality for $x = e$) and for some A and x, $\text{cond}_\infty(A,x) \ll \text{cond}_\infty(A)$ (see, for example, a 3×3 example of Hamming quoted in [17, p. 500]).

Skeel [17], [19] looks in detail at the possible benefits of row scaling for LU factorization. In [17, §4.2] he shows that for the scaling $D^{-1} = \text{diag}(|A||x|)$ the forward error bound is proportional to $\text{cond}_\infty(A,x)$; unfortunately, since x is unknown, this "optimal" scaling is of little practical use. Row equilibration can be regarded as approximating $|x|$ by e in the optimal scaling.

To summarise, we regard row equilibration as a "quick and dirty" way to achieve a "cond-bounded" forward error—quick, because the scaling is trivial to perform, and dirty, because the forward error bound is independent of the right-hand side b and there is no guarantee that a small componentwise backward error will be achieved. In contrast, FPIR produces a small componentwise backward error and has a sharper forward error bound that depends on b (but FPIR may fail to converge).

TABLE 5.1
$A = V_9$, $x = e$, $\kappa_\infty(A) = 4.27e5$, $\mathrm{cond}_\infty(A) = 1.19e3$, $\mathrm{cond}_\infty(A, x) = 1.19e3$.

	$\omega^N(\widehat{y})$	$\omega^R(\widehat{y})$	$\omega^C(\widehat{y})$	$\gamma_\infty(\widehat{y})$
LU with FPIR	2.11e−8	6.25e−7	3.13e−6	1.81e−3
	1.65e−8	1.65e−8	8.26e−8	1.79e−5
LU with equilibration	6.74e−9	1.91e−8	1.72e−7	2.38e−5
QR with FPIR	1.13e−8	7.47e−6	6.73e−5	3.85e−3
	3.51e−9	1.06e−8	8.28e−8	1.44e−5
QR with equilibration	2.3e−8	3.71e−8	3.34e−7	1.60e−4

TABLE 5.2
$A = V_{11}$, $b = e$, $\kappa_\infty(A) = 6.68e7$, $\mathrm{cond}_\infty(A) = 9.17e3$, $\mathrm{cond}_\infty(A, x) = 5.27e1$.

	$\omega^N(\widehat{y})$	$\omega^R(\widehat{y})$	$\omega^C(\widehat{y})$	$\gamma_\infty(\widehat{y})$
LU with FPIR	2.18e−12	2.57e−7	4.82e−6	5.23e−5
	4.57e−12	1.53e−9	5.83e−8	6.83e−7
LU with equilibration	1.22e−10	9.96e−9	2.88e−6	6.24e−5
QR with FPIR	1.95e−11	1.64e−5	1.48e−4	3.59e−4
	4.48e−12	4.86e−9	9.96e−8	1.38e−6
QR with equilibration	3.75e−9	4.75e−9	5.83e−6	1.38e−5

We illustrate our observations with two numerical examples computed using MATLAB in simulated single precision, as in §4. For odd $n = 2k + 1$, let V_n be the Vandermonde matrix with (i, j) element $(-k + j - 1)^{i-1}$. We solved two systems $V_n x = b$ by both LU factorization with partial pivoting and QR factorization, in each case trying both FPIR and the row equilibration discussed above.

The two systems were chosen to illustrate two extreme cases. For the first problem, $V_9 e = b$, reported in Table 5.1, $\mathrm{cond}_\infty(A) = \mathrm{cond}_\infty(A, x) \approx \frac{1}{359}\kappa_\infty(A)$ and row equilibration is about as effective as FPIR as measured by the size of the componentwise backward error and the relative error. For the second system, $V_{11}x = e$, reported in Table 5.2, $\mathrm{cond}_\infty(A, x) \approx \frac{1}{174}\mathrm{cond}_\infty(A) \ll \kappa_\infty(A)$ and FPIR achieves a significantly smaller componentwise backward error and relative error than row equilibration.

We also tried using a scaling obtained by perturbing the equilibrating transformation $D = \mathrm{diag}(|A|e)^{-1}$ to the nearest powers of 2, so as not to introduce rounding errors. This led to final errors sometimes larger and sometimes smaller than with D. In any case, from the point of view of the error bounds the rounding errors introduced by the scaling are easily seen to be insignificant.

REFERENCES

[1] M. ARIOLI, J. W. DEMMEL, AND I. S. DUFF, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.

[2] M. ARIOLI AND A. LARATTA, *Error analysis of an algorithm for solving an underdetermined linear system*, Numer. Math., 46 (1985), pp. 255–268.

[3] F. L. BAUER, *Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme*, Z. Angew. Math. Mech., 46 (1966), pp. 409–421.

[4] A. BJÖRCK, *Stability analysis of the method of seminormal equations for linear least squares problems*, Linear Algebra Appl., 88/89 (1987), pp. 31–48.

[5] R. E. CLINE AND R. J. PLEMMONS, *l_2-solutions to underdetermined systems*, SIAM Rev., 18 (1976), pp. 92–106.

[6] P. E. GILL AND W. MURRAY, *A numerically stable form of the simplex algorithm*, Linear Algebra Appl., 7 (1973), pp. 99–138.

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

[8] W. W. Hager, *Condition estimates*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 311–316.

[9] N. J. Higham, FORTRAN *codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation* (*Algorithm* 674), ACM Trans. Math. Software, 14 (1988), pp. 381–396.

[10] ———, *Algorithm* 694: *A collection of test matrices in* MATLAB, ACM Trans. Math. Software, 17 (1991), pp. 289–305.

[11] ———, *Experience with a matrix norm estimator*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 804–809.

[12] ———, *Iterative refinement enhances the stability of* QR *factorization methods for solving linear equations*, Numer. Anal. Rep. No. 182, Univ. of Manchester, England, 1990.

[13] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice–Hall, Englewood Cliffs, NJ, 1974.

[14] W. Oettli and W. Prager, *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math., 6 (1964), pp. 405–409.

[15] C. C. Paige, *An error analysis of a method for solving matrix equations*, Math. Comp., 27 (1973), pp. 355–359.

[16] M. A. Saunders, *Large-scale linear programming using the Cholesky factorization*, Report CS 252, Computer Science Dept., Stanford Univ., 1972.

[17] R. D. Skeel, *Scaling for numerical stability in Gaussian elimination*, J. Assoc. Comput. Mach., 26 (1979), pp. 494–526.

[18] ———, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., 35 (1980), pp. 817–832.

[19] ———, *Effect of equilibration on residual size for partial pivoting*, SIAM J. Numer. Anal., 18 (1981), pp. 449–454.

[20] P.-Å. Wedin, *Perturbation theory for pseudo-inverses*, BIT, 13 (1973), pp. 217–232.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 15-32, January 1993

002

APPROXIMATIONS OF THE SPECTRAL RADIUS, CORRESPONDING EIGENVECTOR, AND SECOND LARGEST MODULUS OF AN EIGENVALUE FOR SQUARE, NONNEGATIVE, IRREDUCIBLE MATRICES*

ORNA GROSS† AND URIEL G. ROTHBLUM‡

Abstract. Approximations are obtained for the normalized eigenvector of a square, nonnegative, irreducible matrix corresponding to its spectral radius from solutions to linear inequality systems whose feasibility have long been used to characterize lower and/or upper bounds on the spectral radius. These linear inequality systems depend on a parameter that can be viewed as an estimator for the spectral radius. In particular, we derive bounds on the tightness of the resulting approximation to the corresponding eigenvector as a product of a constant (equaling the optimal objective value of a nonlinear, convex optimization problem) and the difference between the spectral radius and its estimator. Bounds are also developed on the second largest modulus of an eigenvalue of a square, nonnegative, irreducible matrix in terms of approximations to its spectral radius and the corresponding normalized eigenvectors. The latter results depend on the methods of Rothblum and Tan [*Linear Algebra Appl.*, 66 (1985), pp. 45-86], who derived bounds on the second largest modulus of an eigenvalue, which depends on the explicit knowledge of the spectral radius and corresponding eigenvector of the underlying matrix.

Key words. spectral radius, nonnegative matrix, eigenvectors, eigenvalues, approximations

AMS(MOS) subject classifications. 15A42, 15A48

1. Introduction.

The well-known Perron–Frobenius theorem asserts that the spectral radius of a square, nonnegative, irreducible matrix is a simple eigenvalue of that matrix with corresponding right and left positive eigenvectors (see § 2 for formal definitions). Furthermore, when the matrix is aperiodic, the powers of the normalized matrix (obtained by dividing the matrix by its spectral radius) have a limit, and the convergence rate of that sequence is geometric with a rate that is determined by the ratio of the second largest modulus of an eigenvalue of the matrix and its spectral radius. Related results hold when the underlying matrix is periodic except that $(C, 1)$ rather than regular limits have to be considered. Thus the spectral radius, corresponding eigenvector, and second largest modulus of the eigenvalues are important characteristics of the matrix. The purpose of this paper is to describe methods for approximating these characteristics.

Nonnegative matrices are useful in describing the dynamics of many multiplicative processes. In such processes, a vectorial input x is transferred, during a single period, into an output vector $x^T P$, where P is a square nonnegative matrix. When such systems are operated over a number of periods, the output at the end of each period is used as the input at the beginning of the next period. For example, finite state Markov chains have the above structure with the input/output vectors representing probability distributions. Also, this structure is shared by input/output production models (Gale [1960]), branching processes (Harris [1963]), Markov reward processes with exponential utility (Howard and Matheson [1972] or Rothblum [1974]), iterative methods in numerical analysis (Varga [1962]), and many other models. When the transition matrix of any one of the aforementioned applications is irreducible, we can use the facts mentioned

* Received by the editors June 21, 1990; accepted for publication (in revised form) February 28, 1991.

† Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel.

‡ Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel (ierurl0@technion) and Rutgers Center for Operations Research, Rutgers University, P.O. Box 5062, New Brunswick, New Jersey 08904 (rothblum@cancer.rutgers.edu).

above about the spectral radius, corresponding eigenvector, and second largest modulus of the eigenvalues to study equilibrium and asymptotic behavior of the underlying system. Furthermore, this can also be accomplished when the transition matrix is reducible by considering the irreducible submatrices corresponding to communicating classes (see Rothblum [1975] and [1981]).

Let P be a square, nonnegative, irreducible matrix with spectral radius ρ. Standard results about nonnegative matrices show that a scalar μ is an upper bound on ρ if and only if the system $Px \leqq \mu x$ has a nonzero, nonnegative solution; and a scalar λ is a lower bound on ρ if and only if the system $Px \geqq \lambda x$ has a nonzero, nonnegative solution (see Theorems 3.1 and 3.3 for details). The main results of the current paper concern bounds on the proximity of normalized solutions of these systems of linear inequalities to the normalized eigenvectors of P corresponding to ρ. Specifically, let f be a (strictly) positive vector and let w be the unique right eigenvector of P corresponding to its spectral radius ρ satisfying the normalization condition $f^T w = 1$. The results we establish in § 3 show that if $\| \ \|$ is an arbitrary norm on R^n and $v \in R^n$ and $\mu \in R$ satisfy

$$Pv \leqq \mu v, \quad v \geqq 0, \quad \text{and} \quad f^T v = 1, \tag{1.1}$$

then $\|v - w\|$ is bounded from above by the product of $(\mu - \rho)$ and the optimal objective value of an optimization problem. The optimization problem does depend on μ, but we show that its objective function is uniformly bounded in μ. Also, we obtain similar results for vectors v in R^n satisfying

$$Pv \geqq \lambda v, \quad v \geqq 0, \quad \text{and} \quad f^T v = 1, \tag{1.2}$$

except that the uniform boundedness of the objective function applies only to scalars λ that are sufficiently close to ρ. Furthermore, we get improved bounds for solutions of the combination of (1.1) and (1.2), i.e., for vectors v in R^n satisfying

$$Pv \leqq \mu v, \quad Pv \geqq \lambda v, \quad v \geqq 0, \quad \text{and} \quad f^T v = 1. \tag{1.3}$$

The above results suggest the use of sequences that approximate the spectral radius ρ of the square, nonnegative, irreducible matrix P to generate sequences of vectors that approximate the corresponding normalized eigenvector w. In particular, if $\{\mu_k\}$ is a sequence of scalars that converges to ρ and $\{v_k\}$ is a sequence of nonnegative vectors satisfying $Pv_k \leqq \mu_k v_k$ and the normalization condition $f^T v_k = 1$, then $\{v_k\}$ converges to w with convergence rate that is bounded by the convergence rate of $\{\mu_k\}$ to ρ. Similarly, if $\{\lambda_k\}$ is a sequence of scalars that converges to ρ and $\{v_k\}$ is a sequence of nonnegative vectors satisfying $Pv_k \geqq \lambda_k v_k$ and the normalization condition $f^T v_k = 1$, then $\{v_k\}$ converges to w with convergence rate that is bounded by the convergence rate of $\{\lambda_k\}$ to ρ. As every scalar is either an upper bound or a lower bound on ρ, we get, from the standard results about nonnegative matrices mentioned at the beginning of the above paragraph, that for every nonnegative scalar β, there exists a nonnegative vector v that satisfies either $Pv \leqq \beta v$ or $Pv \geqq \beta v$. Furthermore, the normalization condition $f^T w = 1$ can be augmented to either system. Thus any sequence of scalars that converges to ρ can be used to generate a sequence of vectors that converges to w. Iterative methods for generating approximating sequences of the spectral radius of a square, nonnegative, irreducible matrix P can be found in Collatz [1942], Hall and Spanier [1968], Marek and Varga [1969], Yamamoto [1966], and Veinott [1971]. Furthermore, Weil [1964], Hamburger, Thompson, and Weil [1967], and Robinson [1974] describe methods for computing the expansion rate for an irreducible von Neumann economic model, a problem that specializes to the problem of determining the spectral radius of an irreducible nonnegative matrix when either the input or the output matrix is the identity.

The derivation of bounds on the tightness of the approximations of the eigenvector of a square, nonnegative, irreducible matrix P with respect to its spectral radius ρ was motivated by the methods of Haviv [1988], who considers the problem of approximating the exact eigenvector of a perturbation of the underlying matrix P. Gross [1989] used such approximations to study the problems considered in the current paper, but the resulting analysis has some serious limitations; see the second remark following Theorem 3.9. We also note that Haviv's approach is to measure the tightness of approximations via coordinate-by-coordinate comparisons and the resulting optimization problems representing the bounds are linear programs. Our approach allows for the use of arbitrary norms with resulting nonlinear convex optimization problems representing the bounds, though we also obtain coordinate-by-coordinate bounds.

Rothblum and Tan [1985] obtained a general method for producing bounds on the second largest modulus of an eigenvalue of a square, nonnegative, irreducible matrix. Their results extended and unified many specific ad hoc bounds that had been obtained earlier. Their methods, as well as all other methods we know of, depend on the explicit knowledge of the spectral radius and corresponding eigenvector of the underlying matrix. This fact creates a problem in numerical implementation because the spectral radius and corresponding eigenvector are frequently not available explicitly, and one must rely on approximations of these quantities. In § 4, we show how the approach of Rothblum and Tan can be augmented to develop bounds on the second largest modulus of an eigenvalue in terms of approximations to its spectral radius and the corresponding (normalized) eigenvector of the given matrix.

As the spectrum of a matrix coincides with that of its scalings, we can apply the above methods for approximating the second largest modulus of an eigenvalue of a given square, nonnegative, irreducible matrix P to any one of its scalings. Generally, this approach results in different bounds than those derived from the matrix P itself; see the examples of Rothblum and Tan [1985] and Gross [1989]. However, we show in § 5 that the resulting approximations and corresponding bounds can be equivalently derived by considering the original matrix and a scaling of the applied norm.

We summarize some notation and preliminary results in § 2 and derive bounds on the tightness of approximations of the eigenvectors in § 3. Next, in § 4, we obtain bounds on the second largest modulus of an eigenvalue using approximations to the spectral radius and corresponding eigenvector. Finally, in § 5, we show that the application of the methods of §§ 3 and 4 to scalings of the given matrix (which is known to have the same spectrum) is equivalent to applying the methods to the matrix itself with a scaling of the underlying norm.

2. Notation and convention. A matrix B is called *nonnegative* or *positive*, written $B \geqq 0$ or $B \gg 0$, if all of its coordinates are nonnegative or positive, respectively. A matrix B is called *semipositive*, written $B > 0$, if $B \geqq 0$ and $B \neq 0$. Given two matrices A and B of the same dimension, we write $A \gg B$ and $B \ll A$, $A \geqq B$ and $B \leqq A$, or $A > B$ and $B < A$ if, respectively, $A - B \gg 0$, $A - B \geqq 0$, or $A - B > 0$. Corresponding definitions and notation apply to vectors.

An $n \times n$, nonnegative matrix B is called *irreducible* if $\Sigma_i B^i \gg 0$, where the summation is taken over $i = 0, \cdots, n$. Such a matrix is called *aperiodic* if $B^t \gg 0$ for some $t = 1, 2, \cdots$.

Let $B \in R^{n \times n}$. The *spectrum* of B will be denoted $\sigma(B)$ and the *spectral radius* of B will be denoted $\rho(B)$, i.e., $\sigma(B)$ is the set of eigenvalues of B and $\rho(B) = \max \{ |\lambda| : \lambda \in \sigma(B)\}$. The celebrated Perron–Frobenius Theorem, e.g., Berman and Plemmons [1979], asserts that the spectral radius of a square, nonnegative matrix is an eigenvalue of that matrix having corresponding semipositive left and right eigenvectors. Furthermore, if the matrix is irreducible, the spectral radius is a simple eigenvalue with corresponding positive left and right eigenvectors. The *second largest modulus of an eigenvalue* of B is denoted $\zeta(B)$, i.e., $\zeta(B) = \max \{ |\lambda| : \lambda \in \sigma(B), \lambda \neq \rho(B) \}$.

A *norm* on R^n is a real-valued function $\| \ \|$ on R^n such that for every x and y in R^n and α in R, $\|x + y\| \leqq \|x\| + \|y\|$, $\|\alpha x\| \leqq |\alpha| \|x\|$, and $\|x\| = 0$ if and only if $x = 0$. For example, for $1 \leqq p < \infty$, the l_p norm on R^n is defined for each $x \in R^n$ by $\|x\|_p = (\Sigma_i |x_i|^p)^{1/p}$, where the summation is taken over $i = 1, \cdots, n$. Also, the l_∞ norm on R^n is defined for each $x \in R^n$ by $\|x\|_\infty = \max \{ |x_i| : i = 1, \cdots, n \}$.

Let $\| \ \|$ be a norm on R^n. The *dual norm* of $\| \ \|$ is denoted $\| \ \|^*$, i.e., for each $a \in R^n$, $\|a\|^* = \max \{ x^T a : x \in R^n, \|x\| \leqq 1 \}$. For $1 \leqq p \leqq \infty$, the dual norm of the l_p norm is known to be the l_q norm, where $q^{-1} + p^{-1} = 1$, in particular, it is computable explicitly. We apply norms and their duals to both column and row vectors without making any notational distinction. Also, we use the notation $\| \ \|$ for the corresponding *matrix norm*, i.e., for $B \in R^{n \times n}$, $\|B\| = \max \{ \|x^T B\| : x \in R^n, \|x\| \leqq 1 \}$ (this definition is convenient for our derivations, though it is nonstandard; the usual definition takes a corresponding maximum over $\|Bx\|$). The *Frobenius norm* of an $n \times n$ matrix B is defined by $\|B\|_F = (\Sigma_{ij} |B_{ij}|^2)^{1/2}$, where the summation is taken over all coordinates of B. Standard results show that $\|B\|_2 \leqq \|B\|_F$, but, $\|B\|_F$ has the advantage of easier computability than $\|B\|_2$ (see, e.g., Rothblum and Tan [1985]).

Given a norm $\| \ \|$ on R^n, a matrix $B \in R^{n \times n}$, and a vector $v \in R^n$, we define the quantity $\tau(B, \| \ \|, v)$ by

$$\tau(B, \| \ \|, v) \equiv \max \{ \|x^T B\| : x \in R^n, \|x\| \leqq 1 \text{ and } x^T v = 0 \}.$$

Closed-form expressions for these functionals when the underlying norm is either the l_1 or the l_∞ norms are given in Rothblum [1984]. Of course, we have $\tau(B, \| \ \|, u) \leqq \|B\|$.

The following result, established in Rothblum and Tan [1985], provides upper bounds on the second largest modulus of an eigenvalue of a square, nonnegative, irreducible matrix.

THEOREM 2.1. *Let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding eigenvector w and let $\| \ \|$ be a norm on R^n. Then, $\zeta(P) \leqq \tau(P, \| \ \|, w)$.*

We note that the conclusion of Theorem 2.1 is trivial if the assumption that $\| \ \|$ is a norm on R^n is tightened to require that it be a complex norm defined on C^n, i.e., $\| \ \|$ maps C^n into R where for every x and y in C^n and α in C, $\|x + y\| \leqq \|x\| + \|y\|$, $\|\alpha x\| \leqq |\alpha| \|x\|$, and $\|x\| = 0$ if and only if $x = 0$. The proof of Theorem 2.1 in Rothblum and Tan [1985] relies on an extension of real norms to complex norms.

Given an $n \times n$ matrix B and subsets K and L of $\{1, \cdots, n\}$, we denote by B_{KL} the *submatrix* of B corresponding to the rows indexed by K and the columns indexed by L. If $K = L$, we write $B_K \equiv B_{KL}$ and call B_K a *principal submatrix* of B. Furthermore, if $K \neq \{1, \cdots, n\}$, we call B_K a *principal strict submatrix* of B. Also, if b is a vector in R^n and K is a subset of $\{1, \cdots, n\}$ we denote by the b_K the corresponding *subvector* of b.

Finally, we find it convenient to call a sequence of vectors *convergent* even if some of the coordinates of the limit are infinite, i.e., $+\infty$ or $-\infty$. Also, if $\{a_k\}$ and $\{b_k\}$ are sequences of real numbers, we write $a_k = O(b_k)$ if a_k / b_k is bounded in k.

3. Approximations of the spectral radius and corresponding eigenvector. In this section we present a number of characterizations of bounds on the spectral radii of square, nonnegative, irreducible matrices, and show how such bounds can be used to obtain approximations of the corresponding normalized eigenvectors. In particular, we describe a bisection algorithm that generates geometrically converging sequences of lower and upper bounds on the spectral radius, and we show how to use such sequences to derive geometrically converging approximations of the corresponding normalized eigenvector. The bisection idea is used in Hamburger, Thompson, and Weil [1967] to compute the expansion rate of a von Neumann economic model; also, Veinott [1971] used the approach explicitly for computing the spectral radius of an irreducible nonnegative matrix.

We start by stating a (standard) characterization of upper bounds on the spectral radius of an $n \times n$ nonnegative, irreducible matrix via feasibility of a linear system having n inequality constraints, one equality constraint, and nonnegative variables; and we obtain (new) bounds on the proximity of solutions of that linear system to the normalized eigenvector of the underlying matrix corresponding to its spectral radius.

THEOREM 3.1. *Let f be a positive vector in R^n and let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding positive eigenvector w. Then a real number μ is an upper bound on ρ if and only if the linear inequality system*

$$Px \leqq \mu x, \quad f^T x = 1, \quad x \geqq 0 \tag{3.1}$$

has a solution. Furthermore, if $\| \; \|$ is any norm on R^n, $x = v$ and μ satisfy (3.1), *and w satisfies the normalization condition $f^T w = 1$, then*

$$\begin{aligned} \|v - w\| \leqq (\mu - \rho) \max \; & \|y\| \\ \text{s.t. } & (P - \mu I)y \leqq g, \\ & f^T y = 0, \\ & y \in R^n, \end{aligned} \tag{3.2}$$

where $g \in R^n$ is the vector defined by $g_i \equiv 1/f_i$, $i = 1, \cdots, n$.

Proof. First, assume that μ is an upper bound on ρ. The positivity of f and w implies that $f^T w > 0$. As $Pw = \rho w \leqq \mu w$, we have that $x = w/f^T w$ satisfies (3.1); thus, (3.1) is feasible. Alternatively, assume that (3.1) is feasible with solution $x = v$. Let u^T be a positive left eigenvector of P corresponding to ρ. Then $u^T v > 0$. Premultiplying the vector-inequality $Pv \leqq \mu v$ by u^T, we conclude that $\rho u^T v = u^T P v \leqq \mu u^T v$, implying that $\rho \leqq \mu$.

Next assume that $\| \; \|$ is a norm on R^n, that $x = v$ and μ satisfy (3.1), and that w satisfies $f^T w = 1$. Then $f^T(v - w) = f^T v - f^T w = 1 - 1 = 0$. Also, our earlier conclusions imply that $\mu \geqq \rho$; hence $(P - \mu I)(v - w) = (P - \mu I)v - (P - \mu I)w \leqq (\mu - \rho)w$, implying that:

$$\begin{aligned} \|v - w\| \leqq \max \; & \|z\| \\ \text{s.t. } & (P - \mu I)z \leqq (\mu - \rho)w, \\ & f^T z = 0, \\ & z \in R^n. \end{aligned} \tag{3.3}$$

The normalization condition $f^T w = 1$ and the positivity of f and w imply that for $i = 1, \cdots, n$, $w_i = f_i w_i / f_i \leqq f^T w / f_i = 1/f_i = g_i$, i.e., $w \leqq g$. As $\mu \geqq \rho$, we conclude that the term $(\mu - \rho)w$ in (3.3) can be replaced by $(\mu - \rho)g$ to obtain the following relaxation of (3.3):

$$\begin{aligned} \|v - w\| \leqq \max \; & \|z\| \\ \text{s.t. } & (P - \mu I)z \leqq (\mu - \rho)g, \\ & f^T z = 0, \\ & z \in R^n. \end{aligned} \tag{3.4}$$

Now, if $\mu \neq \rho$, (3.2) follows directly from (3.4) by applying the change of variable $y = (\mu - \rho)^{-1}z$ and using the positive homogeneity of real norms. Alternatively, assume that $\mu = \rho$ and let v be a vector satisfying (3.1); in particular, $Pv \leqq \mu v = \rho v$. Let u^T be a positive left eigenvector of P corresponding to ρ. As $0 = (\rho u^T - u^T P)v = u^T(\rho v - Pv)$, $u \gg 0$, and $Pv \leqq \mu v = \rho v$, we conclude that $Pv = \rho v$. So, v is an eigenvector of the irreducible matrix P corresponding to its spectral radius ρ, and the normalization conditions $f^T w = f^T v = 1$ imply that $w = v$. Thus (3.2) holds in this case in the form of the trivial inequality $0 \leqq 0$. □

Remarks. 1. Let P be a square, nonnegative, irreducible matrix with spectral radius ρ. Given an upper bound μ on ρ, we can view solutions of (3.1) as approximations of the corresponding normalized eigenvector w, and (3.2) provides us with bounds on the tightness of such approximations. The forthcoming Theorem 3.2 assures that these bounds are always finite.

2. The objective function of the optimization problem occurring in (3.2) is convex and its feasible region is determined by linear inequalities. There are many algorithms for solving nonlinear optimization problems of this class, thus Theorem 3.1 provides us with computable bounds on $\|v - w\|$.

3. The arguments used to deduce (3.2) from (3.4) can be applied to inequality (3.3) to obtain another bound on $\|v - w\|$ given by:

$$(3.5) \qquad \begin{aligned} \|v - w\| \leqq (\mu - \rho) \max\ & \|y\| \\ \text{s.t. } & (P - \mu I)y \leqq w, \\ & f^T y = 0, \\ & y \in R^n. \end{aligned}$$

4. Inequalities (3.3) and (3.5) provide bounds on $\|v - w\|$ that depend on the eigenvector w (and the scalar μ). The dependence on w is an undesirable property because the resulting computation of bounds on the tightness of an approximation to w requires the availability of the unknown vector w. This is why we relaxed the bound given by (3.3) to those given by (3.4) and (3.2), which do not include w.

5. The positivity of f in the characterization of upper bounds on the spectral radius can be replaced by the assumption that $f^T w > 0$. Also, (3.3) and (3.5) hold without the assumption that f is positive, while maintaining the normalization condition $f^T w = 1$. However, the positivity of f is needed for (3.2) and (3.4) to assure that $w \leqq g$ and that no coordinate of g is defined as the reciprocal of zero.

6. The characterization of upper bounds on ρ via feasibility of (3.1) can be modified by replacing the normalization condition $f^T x = 1$ by $\|x\|_2 = 1$. However, as the l_2 norm is not linear, the assertion that $\|v\|_2 = \|w\|_2 = 1$ does not imply $\|v - w\|_2 = 0$, and we do not get a useful analogue of (3.2) by using the l_2 norm. This remark applies to normalization by other norms as well.

We next show that the optimal objective value of the optimization problem on the right-hand side of (3.2) is uniformly bounded for $\mu \geqq \rho$. In particular, this result implies that for every specific selection of μ, the right-hand side of (3.2) is finite. Furthermore, it shows that the bound provided by (3.2) for $\|v - w\|$ depends linearly on the proximity of μ to the spectral radius ρ.

THEOREM 3.2. *Let f be a positive vector in R^n and let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding positive eigenvector w satisfying the normalization condition $f^T w = 1$. Also, let $\| \ \|$ be a norm on R^n. Then the optimization problem on the right-hand side of* (3.2) *is uniformly bounded in $\mu \geqq \rho$.*

Proof. Assume that the optimal objective value of the optimization problem occurring in (3.2) is not uniformly bounded in $\mu \geqq \rho$. Then there exist sequences of scalars $\{\mu_k\}$ and vectors $\{y_k\}$ such that $\lim_{k\to\infty} \|y_k\| = \infty$ and for each $k = 1, 2, \cdots$,

$$(P - \mu_k I)y_k \leqq g, \quad f^T y_k = 0, \quad \text{and} \quad \mu_k \geqq \rho.$$

The sequence $\{(\mu_k, y_k/\|y_k\|)\}$ has a convergent subsequence. Let (μ, d) be the limit of such a subsequence. In particular, d is a finite, nonzero vector. Further, by dividing the inequalities $f^T y_k = 0$ by $\|y_k\|$, we get from continuity arguments that $f^T d = 0$. We continue our analysis by distinguishing between the case where μ is finite and the case where it is not.

We first consider the case where μ is finite. Evidently, $\mu \geqq \rho$. Also, dividing the inequalities $(P - \mu_k I)y_k \leqq g$ by $\|y_k\|$, we get from continuity arguments that $(P - \mu I)d \leqq 0$. Let $K \equiv \{i = 1, \cdots, n : d_i < 0\}$ and $L \equiv \{i = 1, \cdots, n : d_i \geqq 0\}$. As $f \gg 0$, $d \neq 0$, and $f^T d = 0$, we have that K and L are nonempty and differ from $\{1, \cdots, n\}$. Let $\rho(P_K)$ be the spectral radius of the submatrix P_K of P and let u^T be a corresponding semipositive left eigenvector. In particular, as P_K is a nonempty, principal, strict submatrix of P, we have that $\rho(P_K) < \rho$ (see Varga [1962, p. 30]). Also, as $d_K \ll 0$, we have that $u^T d_K < 0$, and as $Pd \leqq \mu d$,

$$P_K d_K \leqq P_K d_K + P_{KL} d_L = (Pd)_K \leqq \mu d_K.$$

Upon premultiplying the last inequality by u^T, we conclude that $\rho(P_K) u^T d_K = u^T P_K d_K \leqq \mu u^T d_K$. As $u^T d_K < 0$, we get a contradiction to the fact that $\rho(P_K) < \rho$.

We next consider the case where μ is not finite. As the sequence $\{\mu_k\}$ is bounded from below by ρ, we have that $\mu = \infty$. Dividing the inequalities $(P - \mu_k I)y_k \leqq g$ by $\mu_k \|y_k\|$, we get from continuity arguments that $-d \leqq 0$. As $d \neq 0$ and $f^T d = 0$, we obtain a contradiction to the assumption that f is positive.

Either of the two cases we considered led to contradiction. Hence we conclude that the optimal objective value of the optimization problem occurring in (3.2) is, indeed, uniformly bounded for $\mu \geqq \rho$. □

We next modify Theorems 3.1 and 3.2 by considering lower bounds on the spectral radius rather than upper bounds.

THEOREM 3.3. *Let f be a positive vector in R^n, let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding positive eigenvector w. Then a real number λ is a lower bound on ρ if and only if the linear inequality system*

$$Px \geqq \lambda x, \quad f^T x = 1, \quad x \geqq 0 \tag{3.6}$$

has a solution. Furthermore, if $\|\ \|$ is any norm on R^n, $x = v$ and λ satisfy (3.6), *and w satisfies the normalization condition $f^T w = 1$, then*

$$\begin{aligned} \mathrm{dv}v - w\| \leqq (\rho - \lambda) \max\ & \|y\| \\ \text{s.t. } & (P - \lambda I)y \geqq -g, \\ & f^T y = 0, \\ & y \in R^n, \end{aligned} \tag{3.7}$$

where $g \in R^n$ is the vector defined by $g_i \equiv 1/f_i$, $i = 1, \cdots, n$.

Proof. The characterization of lower bounds on ρ via feasibility of (3.6) follows from the arguments used to establish the corresponding characterization for upper bounds in Theorem 3.1. Also, the arguments used in the proof of Theorem 3.1 with reversed inequalities and λ replacing μ show that if $\| \ \|$ is any norm on R^n, $x = v$ and λ satisfy (3.6), and w satisfies the normalization condition $f^T w = 1$, then

$$\begin{aligned} \|v - w\| &\leqq \max \|z\| \\ &\text{s.t. } (P - \lambda I)z \geqq (\lambda - \rho)w = -(\rho - \lambda)w, \\ &\quad f^T z = 0, \\ &\quad z \in R^n, \end{aligned} \tag{3.8}$$

and that the term $-(\rho - \lambda)w$ in the above bound can be replaced by $-(\rho - \lambda)g$. Now, if $\lambda \neq \rho$, (3.7) follows from (3.8) by applying the change of variable $y = (\rho - \lambda)^{-1}z$ and using the positive homogeneity of real norms. Alternatively, if $\lambda = \rho$, (3.7) follows from the corresponding arguments used in the proof of Theorem 3.1. □

The remarks following Theorem 3.1 apply to Theorem 3.3. In particular, (3.7) provides computable bounds on $\|v - w\|$. Also, (3.8) can be transformed into a variant corresponding to (3.5). However, the resulting inequality, as well as the original (3.8), are not as useful as (3.7) because they provide bounds on the tightness of an approximation to w which depend on w itself. Finally, the positivity of f is needed only for the definition of g as a finite vector and for the inequality $w \leqq g$.

Unfortunately, the optimization problem occurring in (3.7) is not necessarily bounded, as demonstrated by the following example. Let

$$P \equiv \begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}, \quad f \equiv \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \lambda \equiv 1, \quad \text{and} \quad y^s \equiv \begin{pmatrix} s \\ -s \end{pmatrix} \quad \text{for} \quad s \geqq 0.$$

Then, for every $s \geqq 0$, $f^T y^s = 0$ and $(P - \lambda I)y^s = (0, 2s)^T \geqq -(1, 1)^T = -g$, while $\lim_{s \to \infty} \|y^s\|_\infty = \infty$.

The next result shows that, despite the above example, the right-hand side of (3.7) is uniformly bounded for λ which is sufficiently close to the spectral radius ρ.

THEOREM 3.4. *Let f be a positive vector in R^n and let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding positive eigenvector w satisfying the normalization condition $f^T w = 1$. Also, let $\| \ \|$ be a norm on R^n. Then for some $\lambda^* < \rho$, the optimization problem on the right-hand side of* (3.7) *is uniformly bounded in λ satisfying $\lambda^* \leqq \lambda \leqq \rho$.*

Proof. Assume that the conclusion of our theorem is false. Then there exist sequences of scalars $\{\lambda_k\}$ and vectors $\{y_k\}$ such that $\lim_{k \to \infty} \|y_k\| = \infty$, $\lim_{k \to \infty} \lambda_k = \rho$, and for each $k = 1, 2, \cdots$

$$(P - \lambda_k I)y_k \geqq -g, \quad f^T y_k = 0, \quad \text{and} \quad \lambda_k \leqq \rho.$$

The sequence $\{y_k / \|y_k\|\}$ has a convergent subsequence. Let d be the limit of such a subsequence. In particular, d is a finite, nonzero vector. Furthermore, by dividing the inequalities $(P - \lambda_k I)y_k \geqq -g$ and $f^T y_k = 0$ by $\|y_k\|$, we get from continuity arguments that $(P - \rho I)d \geqq 0$ and $f^T d = 0$. Let $K \equiv \{i = 1, \cdots, n : d_i > 0\}$ and $L \equiv \{i = 1, \cdots, n : d_i \leqq 0\}$. As $f \gg 0$, $d \neq 0$ and $f^T d = 0$, we have that K and L are nonempty and differ from $\{1, \cdots, n\}$. Let $\rho(P_K)$ be the spectral radius of the submatrix P_K of P and let u^T be a corresponding semipositive left eigenvector. In particular, as P_K is a nonempty principal strict submatrix of P, we have that $\rho(P_K) < \rho$ (see Varga [1962, p. 30]). Also, as $d_K \gg 0$, we have $u^T d_K > 0$, and as $Pd \geqq \rho d$,

$$P_K d_K \geqq P_K d_K + P_{KL} d_L = (Pd)_K \geqq \rho d_K.$$

Upon premultiplying the last inequality by u^T, we get that $\rho(P_K)u^T d_K = u^T P_K d_K \geqq \rho u^T d_K$. As $u^T d_K > 0$, we obtain a contradiction to the fact that $\rho(P_K) < \rho$. Hence we conclude that the optimal objective value of the optimization problem occurring in (3.7) is, indeed, uniformly bounded for sufficiently large $\lambda \leqq \rho$. □

Reconsider the example preceding Theorem 3.4 with the norm $\| \ \|_\infty$. Then the optimization problem occurring in (3.7) has the form

$$\begin{aligned} &\max \|y\|_{\infty 1} \\ &\text{s.t.}\ (2-\lambda)y_1 + y_2 \geqq -1, \\ &\qquad 2y_1 + (1-\lambda)y_2 \geqq -1, \\ &\qquad y_1 + y_2 = 0. \end{aligned}$$

Using the substitution $y_2 = -y_1$, the above problem reduces to

$$\begin{aligned} &\max |y_1| \\ &\text{s.t.}\ (1-\lambda)y_1 \geqq -1, \\ &\qquad (1+\lambda)y_1 \geqq -1. \end{aligned}$$

This problem is unbounded for $\lambda = 1$ (as we have already observed) and is bounded for $1 < \lambda \leqq 3$ with optimal solution $y_1 = (\lambda - 1)^{-1}$. In particular, for every λ^* satisfying $1 < \lambda^* \leqq 3$, the above (parametric) problem is uniformly bounded for $\lambda^* \leqq \lambda \leqq 3$.

Let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and let λ and μ be, respectively, lower and upper bounds on ρ. Then the interval $[\lambda, \mu]$ contains ρ. The results of either Theorem 3.1 or Theorem 3.3 can next be used to produce a smaller interval which maintains the property of containing ρ. Specifically, suppose f is a positive vector in R^n and β is a scalar satisfying $\lambda < \beta < \mu$. Then $\lambda \leqq \rho \leqq \beta$ if and only if the linear system with n nonnegative variables, n inequalities, and one equation,

$$(3.9) \qquad Px \leqq \beta x, \quad f^T x = 1, \quad x \geqq 0,$$

has a solution. Also, $\beta \leqq \rho \leqq \mu$ if and only if the linear system

$$(3.10) \qquad Px \geqq \beta x, \quad f^T x = 1, \quad x \geqq 0$$

has a solution. Thus, by checking feasibility/infeasibility of either (3.9) or (3.10), we can identify a subinterval of $[\lambda, \mu]$, which maintains the property of containing ρ. By selecting $\beta \equiv 2^{-1}(\lambda + \mu)$, either of the two approaches bisects the interval $[\lambda, \mu]$ and produces an interval with half the length of $[\lambda, \mu]$.

It is easy to obtain initial lower and upper bounds on the spectral radius. Specifically, let v be any positive vector. Then for $\mu \equiv \max\{(Pv)_i/v_i : i = 1, \cdots, n\}$ and $\lambda \equiv \min\{(Pv)_i/v_i : i = 1, \cdots, n\}$, $\lambda v \leqq Pv \leqq \mu v$; hence Theorems 3.1 and 3.3 imply that $\lambda \leqq \rho \leqq \mu$. So, lower and upper bounds on ρ are computable from any positive vector. Of course, zero can replace λ as a lower bound on ρ. Successive applications of the bisection procedure described in the above paragraph produce a sequence of upper bounds $\{\mu_k\}$ and a sequence of lower bounds $\{\lambda_k\}$ on ρ, such that $(\mu_k - \lambda_k) = (\mu - \lambda)2^{-k}$. In particular, the selection of any point ρ_k from the interval $[\lambda_k, \mu_k]$, for $k = 1, 2, \cdots$, will generate a sequence with $|\rho_k - \rho| = O(2^{-k})$. We refer to the resulting algorithm as the *bisection algorithm*. This approach is used in Hamburger, Thompson, and Weil [1967] to compute the expansion rate of a von Neumann economic model—a problem that generalizes the one we consider here of computing the spectral radius. But our evaluation of whether or not $\beta \leqq \rho$ requires the solution of a system of linear inequalities, whereas Hamburger, Thompson, and Weil compute the value of a two-person zero-sum game in matrix form.

Theorems 3.1–3.4 show that sequences of approximations of the spectral radius can be used to generate sequences of approximations of the corresponding normalized eigenvector with convergence rates that are dominated by those of the estimators of the spectral radius. Specifically, we obtain the following corollaries.

COROLLARY 3.5. *Let f be a positive vector in R^n and let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding positive eigenvector w satisfying the normalization condition $f^Tw = 1$. Also, let $\{\mu_k\}$ be a sequence of upper bounds on ρ with* $\lim_{k\to\infty} \mu_k = \rho$ *and let $\{v_k\}$ be a sequence of vectors such that for each $k = 1, 2, \cdots$, $x = v_k$ satisfies* (3.1) *with $\mu = \mu_k$. Then, $\|v_k - w\| = O(\mu_k - \rho)$.*

COROLLARY 3.6. *Let f be a positive vector in R^n and let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding positive eigenvector w satisfying the normalization condition $f^Tw = 1$. Also, let $\{\lambda_k\}$ be a sequence of lower bounds on ρ with* $\lim_{k\to\infty} \lambda_k = \rho$ *and let $\{v_k\}$ be a sequence of vectors such that for each $k = 1, 2, \cdots$, $x = v_k$ satisfies* (3.6) *with $\lambda = \lambda_k$. Then $\|v_k - w\| = O(\rho - \lambda_k)$.*

Let f be a positive vector in R^n and let P be a square, nonnegative, irreducible matrix with spectral radius ρ and corresponding positive eigenvector w satisfying the normalization condition $f^Tw = 1$. The bisection algorithm generates a sequence of lower bounds $\{\lambda_k\}$ and a sequence of upper bounds $\{\mu_k\}$ on ρ such that $|\rho - \mu_k| = O(2^{-k})$ and $|\rho - \lambda_k| = O(2^{-k})$. By finding solutions v_k to either (3.1) with $\mu = \mu_k$ or to (3.6) with $\lambda = \lambda_k$, for $k = 1, 2, \cdots$, we get from Corollary 3.5 or Corollary 3.6, respectively, that $\|v_k - w\| = O(2^{-k})$. Thus, the sequence $\{v_k\}$ is a geometrically converging sequence of approximations of w.

We next combine the characterizations of upper and lower bounds on the spectral radius given in Theorems 3.1 and 3.3, and use two-sided bounds to obtain approximations of the corresponding normalized eigenvector. The bounds we get on the tightness of the resulting approximations improve on those obtained in Theorems 3.1 and 3.3.

THEOREM 3.7. *Let f be a positive vector in R^n and let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding positive eigenvector w. Then real numbers λ and μ are, respectively, lower and upper bounds on ρ if and only if the linear inequality system*

$$(3.11) \qquad Px \leqq \mu x, \quad Px \geqq \lambda x, \quad f^Tx = 1, \quad x \geqq 0$$

is feasible. Furthermore, if $\|\ \|$ is any norm on R^n, if $x = v$, if μ and λ satisfy (3.11), *and if w satisfies the normalization condition $f^Tw = 1$, then*

$$\|v - w\| \leqq (\mu - \lambda) \max \|y\|$$

$$(3.12) \qquad \begin{aligned} \text{s.t. } &(P - \lambda I)y \leqq v, \\ &(P - \mu I)y \geqq -v, \\ &f^Ty = 0, \\ &y \in R^n. \end{aligned}$$

Proof. If λ and μ are, respectively, lower and upper bounds on ρ, then w/f^Tw satisfies (3.11). Alternatively, if (3.11) is feasible, so are (3.1) and (3.6), and Theorems 3.1 and 3.3 imply that λ and μ are, respectively, lower and upper bounds on ρ.

Next assume that $\|\ \|$ is a norm on R^n; that $x = v$, λ, and μ satisfy (3.11); and that w satisfies the normalization condition $f^Tw = 1$. Then $f^T(v - w) = f^Tv - f^Tw = 1 - 1 = 0$. Also, our earlier conclusions show that $\lambda \leqq \rho \leqq \mu$; hence

$$(P - \lambda I)(v - w) = (P - \lambda I)v - (P - \lambda I)w \leqq (\mu - \lambda)v - (\rho - \lambda)w \leqq (\mu - \lambda)v$$

and

$$(P - \mu I)(v - w) = (P - \mu I)v - (P - \mu I)w \geqq (\lambda - \mu)v - (\rho - \mu)w \geqq -(\mu - \lambda)v,$$

implying that

$$\|v - w\| \leqq \max \|z\|$$

$$\text{(3.13)} \qquad \begin{aligned} \text{s.t. } &(P - \lambda I)z \leqq (\mu - \lambda)v,\\ &(P - \mu I)z \geqq -(\mu - \lambda)v,\\ &f^T z = 0,\\ &z \in R^n. \end{aligned}$$

Now, if $\mu \neq \lambda$, (3.12) follows directly from (3.13) by applying the change of variable $y = (\mu - \lambda)^{-1}z$ and using the positive homogeneity of real norms. Alternatively, assume that $\mu = \lambda$. Let v be a vector satisfying (3.11). Then $Pv \leqq \mu v = \rho v = \lambda v \leqq Pv$. So, $Pv = \rho v$, i.e., v is an eigenvector of P corresponding to ρ. The normalization conditions $f^T w = f^T v = 1$ next imply that $w = v$. Thus (3.12) holds in this case in the form of the trivial inequality $0 \leqq 0$. □

Remarks. 1. Let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ. Given an upper bound μ and lower bound λ on ρ, we can view solutions of (3.11) as approximations of the corresponding normalized eigenvector w, and (3.12) provides us with bounds on the tightness of such approximations. The forthcoming Theorem 3.8 assures that these bounds are always finite.

2. The bounds on $\|v - w\|$ given by (3.12) are computable. Furthermore, they require the knowledge of v, μ, and λ, but they do not depend on the eigenvector w, which is the target of our approximation efforts. See remarks 2, 3, and 4 following Theorem 3.1.

3. The positivity of f is required neither for the characterization of pairs of upper and lower bounds on the spectral radius, nor for the derivation of inequality (3.12); see remark 5 following Theorem 3.1. However, the positivity of f is required for our proof of the finiteness of the right-hand side of (3.12) given in the forthcoming Theorem 3.8.

4. The characterization of upper bounds on ρ via feasibility of (3.11) can be modified by replacing the normalization condition $f^T x = 1$ by $\|x\|_2 = 1$. But this approach cannot be extended to get (3.12); see remark 6 following Theorem 3.1.

5. Suppose that $f_k = \min \{f_i : i = 1, \cdots, n\}$. Then $v^T f \geqq f_k \|v\|_1$ and

$$\|g\|_\infty = g_k = 1/f_k = v^T f / f_k \geqq \|v\|_1.$$

We conclude that the average coordinate of the vector v is smaller than or equal to the maximal coordinate of g divided by n. This observation suggests that the optimal objective of the optimization problem occurring in (3.12) is smaller than that of the one in either (3.2) or (3.7) by a factor of n.

We next show that the optimal objective function of the optimization problem occurring in (3.12) is uniformly bounded in v, λ, and μ satisfying (3.11).

THEOREM 3.8. *Let f be a positive vector in R^n and let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding positive eigenvector w satisfying the normalization condition $f^T w = 1$. Also, let $\| \ \|$ be a norm on R^n. Then the optimization problem on the right-hand side of* (3.12) *is uniformly bounded in v, λ, and μ where $x = v$, λ, and μ satisfy* (3.11).

Proof. Assume that the optimal objective value of the optimization problem occurring in (3.12) is not uniformly bounded in v, λ, and μ where $x = v$, λ, and μ satisfy

(3.11). Then there exist sequences of scalars $\{\mu_k\}$ and $\{\lambda_k\}$ and sequences of vectors $\{v_k\}$ and $\{y_k\}$ such that $\lim_{k\to\infty} \|y_k\| = \infty$ and for each $k = 1, 2, \cdots$

$$Pv_k \leqq \mu_k v_k, \quad Pv_k \geqq \lambda_k v_k, \quad f^T v_k = 1, \quad v_k \geqq 0,$$

and

$$(P - \lambda_k I)y_k \leqq v_k, \quad (P - \mu_k I)y_k \geqq -v_k, \quad f^T y_k = 0.$$

As $f^T v_k = 1$ and $v_k \geqq 0$ for each k and as f is positive, we have that the sequence $\{v_k\}$ is bounded. Also, the first part of Theorem 3.7 assures that for each k, $\lambda_k \leqq \rho \leqq \mu_k$.

The sequence $\{(\lambda_k, \mu_k, y_k)/\|y_k\|\}$ has a convergent subsequence. Let (λ, μ, d) be the limit of such a subsequence. In particular, d is a finite, nonzero vector and $\lambda \leqq \rho \leqq \mu$. Furthermore by dividing the inequalities $f^T y_k = 0$ by $\|y_k\|$, we get from continuity arguments that $f^T d = 0$. We continue our analysis by distinguishing between the case where μ and λ are finite and the case where either is infinite.

We first consider the case where both μ and λ are finite. Dividing the inequalities $(P - \lambda_k I)y_k \leqq v_k$ and $(P - \mu_k I)y_k \geqq -v_k$ by $\|y_k\|$, we get from continuity arguments and the boundedness of the sequence $\{v_k\}$ that $\mu d \leqq Pd \leqq \lambda d$; in particular, $(\mu - \lambda)d \leqq 0$. Also, as $\lambda \leqq \rho \leqq \mu$ we have that $(\mu - \lambda) \geqq 0$, and as f is positive and $f^T d = 0$ it is not true that $d \leqq 0$. It follows from these observations that $\mu = \lambda = \rho$, implying that $Pd = \rho d$. So, d is an eigenvector of P corresponding to ρ. As $f^T d = 0$ it now follows that $d = 0$, a contradiction.

We next consider the case where either μ or λ are not finite. If μ is not finite, then $\mu = \infty$, as the sequence $\{\mu_k\}$ is bounded from below by ρ. Dividing the inequalities $(P - \mu_k I)y_k \geqq -v_k$ by $\mu_k\|y_k\|$ we get from continuity arguments and the boundedness of the sequence $\{v_k\}$ that $-d \geqq 0$. As $f \gg 0$, $d \neq 0$, and $f^T d = 0$, we conclude that $d = 0$, another contradiction. A similar contradiction is obtained if λ is not finite. □

In any practical use of our development, we consider only nonnegative lower bounds on the spectral radius, in which case the proof of Theorem 3.8 simplifies somehow as the case where λ is not finite is redundant.

Corollaries 3.5 and 3.6 show that if $\{\mu_k\}$ and $\{\lambda_k\}$ are sequences of upper and lower bounds on ρ, and if for each k, v_k is a solution of (3.11) with $\mu = \mu_k$ and $\lambda = \lambda_k$, then $\|v_k - w\| = O(\mu_k - \lambda_k)$. Consequently, if $\{\mu_k\}$ and $\{\lambda_k\}$ are generated by the bisection algorithm, we get that the resulting sequence $\{v_k\}$ has $\|v_k - w\| = O(2^{-k})$.

The arguments used to establish the bounds on $\|v - w\|$ given in Theorems 3.1, 3.3, and 3.7 can be used to obtain bounds on the individual coordinates of $v - w$. For example, under the assumptions of the above theorems, respectively, lower and upper bounds on the coordinates $(v - w)_i$, $i = 1, \cdots, n$, can be obtained by considering either of the following pairs of linear programs:

$$(3.14) \qquad \begin{array}{ll} (\mu - \rho)\min y_i & (\mu - \rho)\max y_i \\ \quad \text{s.t. } (P - \mu I)y \leqq g, & \quad \text{s.t. } (P - \mu I)y \leqq g \\ \quad f^T y = 0, & \quad f^T y = 0 \\ \quad y \in R^n & \quad y \in R^n, \end{array}$$

$$(3.15) \qquad \begin{array}{ll} (\rho - \lambda)\min y_i & (\rho - \lambda)\max y_i \\ \quad \text{s.t. } (P - \lambda I)y \geqq -g, & \quad \text{s.t. } (P - \lambda I)y \geqq -g, \\ \quad f^T y = 0, & \quad f^T y = 0, \\ \quad y \in R^n, & \quad y \in R^n, \end{array}$$

and

$$(3.16)\qquad \begin{array}{ll} (\mu-\lambda)\min y_i & (\mu-\lambda)\max y_i \\ \text{s.t. } (P-\lambda I)y \leqq v, & \text{s.t. } (P-\lambda I)y \leqq v, \\ (P-\mu I)y \geqq -v, & (P-\mu I)y \geqq -v, \\ f^T y = 0, & f^T y = 0, \\ y \in R^n, & y \in R^n. \end{array}$$

We note that the arguments used to establish Theorems 3.2, 3.4, and 3.8 show that the above optimization problems have (uniformly) bounded objectives over the corresponding ranges of the parameters.

The use of linear programs like those in (3.14)–(3.16) to obtain bounds on the individual coordinates of $v - w$, where w is an eigenvector corresponding to the spectral radius ρ and v is its approximation, was first derived by Haviv [1988]. Haviv's approach was to consider the dual programs, rather than rely on the homogeneity of the optimal objective of a linear program with the right-hand side.

Theorems 3.1, 3.3, and 3.7 show that approximations to the spectral radius of a square, nonnegative, irreducible matrix yield approximations to the corresponding eigenvector. We next show a converse of this result, where vectors that can be viewed as approximate eigenvectors yield approximations of the spectral radius.

THEOREM 3.9. *Let f be a positive vector in R^n and let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding positive eigenvector w satisfying the normalization condition $f^Tw = 1$. Furthermore, let v be a positive vector in R^n with $f^Tv = 1$, let β be a positive number, and let $\delta \equiv Pv - \beta v$. Then*

$$(3.17)\qquad |\rho - \beta| \leqq \max\{|\delta_i|/v_i : i = 1, \cdots, n\}.$$

Furthermore, if $\delta \geqq 0$, then $\rho \geqq \beta$; and if $\delta \leqq 0$, then $\rho \leqq \beta$.

Proof. Let u be a positive left eigenvector of P corresponding to its spectral radius ρ. Then

$$(\rho - \beta)u^Tv = u^T(Pv - \beta v) = u^T\delta.$$

Thus $|\rho - \beta| = |u^T\delta|/u^Tv$. Let $K \equiv \max\{|\delta_i|/v_i : i = 1, \cdots, n\}$. Then (3.17) follows from the inequalities

$$\begin{aligned} |u^T\delta|/u^Tv &\leqq \left(\sum_{i=1}^n u_i|\delta_i|\right)\bigg/\left(\sum_{i=1}^n u_iv_i\right) \\ &\leqq \left(\sum_{i=1}^n u_ikv_i\right)\bigg/\left(\sum_{i=1}^n u_iv_i\right) = k. \end{aligned}$$

The remaining conclusions of the theorem follow from Theorems 3.1 and 3.3. □

Remarks. 1. In most computations of the spectral radius and corresponding eigenvector of a square, nonnegative, irreducible matrix, we end up with a scalar β and a positive (normalized) vector v such that $\delta \equiv Pv - \beta v$ is not the zero vector, but is a small vector with respect to any given norm. So, we can view such a vector v as an approximate eigenvector of P with respect to β. In this case, (3.17) provides a bound on the proximity of β to the spectral radius. Furthermore, for

$$\mu = \beta + \max\{\delta_i/v_i : i = 1, \cdots, n\} \quad\text{and}\quad \lambda = \beta + \min\{\delta_i/v_i : i = 1, \cdots, n\},$$

$Pv \leqq \mu v$, $Pv \geqq \lambda v$, and $|\mu - \lambda| \leqq 2|||\delta|||_\infty$, for the norm $||| \; |||_\infty$ defined for $x \in R^n$ by $|||x|||_\infty \equiv \max\{x_i/v_i : i = 1, \cdots, n\}$. So, Theorems 3.1, 3.3, and 3.7 provide bounds on $|||v - w|||_\infty$ in terms of a product of $|||\delta|||_\infty$, and a computable constant (which depends, respectively, on μ, on λ, or on μ, λ, and v).

2. Under the assumptions of Theorem 3.9, $(P - \delta f^T)v = Pv - \delta = \beta v$; hence v is an eigenvector of $P - \delta f^T$ with respect to the eigenvalue β. If the matrix $P - \delta f^T$ is nonnegative, we conclude that β is its spectral radius, a fact that can be used to derive approximations of the spectral radius of P and of the corresponding normalized eigenvector (see Haviv [1988] and Gross [1989]). But the nonnegativity of $P - \delta f^T$ is a strong and restrictive assumption, which we do not wish to impose.

4. Upper bounds on the second largest modulus of an eigenvalue via approximations of the spectral radius and corresponding eigenvector. Theorem 2.1 gives bounds on the second largest modulus of an eigenvalue of a square, nonnegative, irreducible matrix in terms of the spectral radius of the matrix and corresponding eigenvector. In the current section, we develop bounds that depend on approximations of these parameters.

We start by showing that the functional $\tau(.,.,.)$, defined in § 2 and used in Theorem 2.1, is Lipschitz continuous in its third variable over hyperplanes.

LEMMA 4.1. *Let $B \in R^{n \times n}$, $f \in R^n$, and let $\| \; \|$ be a norm on R^n. Then for every pair of vectors u and v in R^n satisfying $f^T u = f^T v = 1$,*

$$\begin{aligned} \tau(B, \| \; \|, u) &\leqq \tau(B, \| \; \|, v)(1 + \|u - v\|^* \|f\|) + \|u - v\|^* \|f^T B\| \\ &\leqq \tau(B, \| \; \|, v) + \|u - v\|^* (\|f\| \|B\| + \|f^T B\|) \qquad (4.1) \\ &\leqq \tau(B, \| \; \|, v) + 2\|u - v\|^* \|f\| \|B\|. \end{aligned}$$

Proof. Compactness arguments assure that $\tau(B, \| \; \|, u) = \|y^T B\|$ for some $y \in R^n$ with $\|y\| \leqq 1$ and $y^T u = 0$. Now if $y^T B = 0$, then $\tau(B, \| \; \|, u) = 0$ and the first inequality of (4.1) is trivial. Also, if $y - (y^T v)f = 0$, then $0 = [y - (y^T v)f]^T u = y^T u - (y^T v) \times (f^T u) = -(y^T v)$. So, y is feasible for the optimization problem defining $\tau(B, \| \; \|, v)$, implying that $\tau(B, \| \; \|, v) \geqq \|y^T B\| = \tau(B, \| \; \|, u)$ and, again, the first inequality of (4.1) follows. Next consider the case where $y^T B \neq 0$ and $y - (y^T v)f \neq 0$. The assumption that $y^T B \neq 0$ and the optimality of y for the optimization problem defining $\tau(B, \| \; \|, u)$ imply that $\|y\| = 1$. Also, the assumption that $y - (y^T v)f \neq 0$ implies that the real vector

$$z \equiv [y - (y^T v)f] / \|y - (y^T v)f\|$$

is well defined. In particular, $\|z\| = 1$ and $z^T v = 0$, i.e., z is feasible for the optimization problem defining $\tau(B, \| \; \|, v)$. Thus, as $y^T u = 0$ and $\|y\| = 1$,

$$\begin{aligned} \tau(B, \| \; \|, v) &\geqq \|z^T B\| = \|y^T B - \{[y^T(v - u)]f^T B\}\| \|y - [y^T(v - u)]f\|^{-1} \\ &\geqq \|y^T B - \{[y^T(v - u)]f^T B\}\| \{\|y\| + \|[y^T(v - u)]f\|\}^{-1} \\ &\geqq \{\|y^T B\| - \|[y^T(v - u)]f^T B\|\}\{\|y\| + \|[y^T(v - u)]f\|\}^{-1} \\ &\geqq \{\tau(B, \| \; \|, u) - \|y\| \|v - u\|^* \|f^T B\|\}\{\|y\| + \|y\| \|v - u\|^* \|f\|\}^{-1} \\ &= \{\tau(B, \| \; \|, u) - \|v - u\|^* \|f^T B\|\}\{1 + \|v - u\|^* \|f\|\}^{-1}, \end{aligned}$$

implying that

$$\tau(B, \| \; \|, u) \leqq \tau(B, \| \; \|, v)\{1 + \|v - u\|^* \|f\|\} + \|v - u\|^* \|f^T B\|.$$

So, the first inequality of (4.1) has been established. The remaining inequalities of (4.1) follow directly from the standard inequalities $\tau(B, \| \; \|, v) \leqq \|B\|$ and $\|f^T B\| \leqq \|f\| \|B\|$. □

We note that the first bound in (4.1) is the tightest one, but the other two are included because they are useful in our forthcoming development and because of their simplicity.

COROLLARY 4.2. *Let $B \in R^{n \times n}$, $f \in R^n$, and $\| \ \|$ be a norm on R^n. Then for every pair of vectors u and v in R^n satisfying $f^T u = f^T v = 1$,*

$$(4.2) \quad \begin{aligned} |\tau(B, \| \ \|, u) - \tau(B, \| \ \|, v)| &\leqq \|u - v\|^*(\|f\| \|B\| + \|f^T B\|) \\ &\leqq 2\|u - v\|^* \|f\| \|B\|. \end{aligned}$$

In particular, $\tau(B, \| \ \|, .)$ is Lipschitz continuous on $\{x \in R^n : x^T f = 1\}$ (which implies that $\tau(B, \| \ \|, .)$ is continuous on that set).

THEOREM 4.3. *Let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding eigenvector w; let $f \in R^n$ satisfy $f^T w = 1$; and let $\| \ \|$ be a norm on R^n. Then for each $v \in R^n$ satisfying $f^T v = 1$,*

$$(4.3) \quad \begin{aligned} \zeta(P) &\leqq \tau(P, \| \ \|, v)(1 + \|v - w\|^* \|f\|) + \|v - w\|^* \|f^T P\| \\ &\leqq \tau(P, \| \ \|, v) + \|v - w\|^*(\|f\| \|P\| + \|f^T P\|) \\ &\leqq \tau(P, \| \ \|, v) + 2\|v - w\|^* \|f\| \|P\|. \end{aligned}$$

Proof. By Theorem 2.1, $\zeta(P) \leqq \tau(P, \| \ \|, w)$. Combining this inequality with (4.1) immediately yields (4.3). □

Theorem 4.3 provides bounds on $\zeta(P)$ in terms of an approximation v to the normalized eigenvector w of the underlying matrix P with respect to its spectral radius ρ. As the bound contains the term $\|v - w\|^*$, we can use the results of § 3 to derive bounds that do not contain w. Specifically, given a positive vector v which satisfies (3.1), (3.6), or (3.11) for scalars μ and/or λ, respectively, we have that $\|v - w\|^*$ is bounded by the product of $|\mu - \lambda|$ and a (computable) constant, which is uniformly bounded over the corresponding range. In particular, these constants are the optimal objective values of an optimization problem involving the dual norm. Of course, explicit expressions of the dual norm are available for the l_p norms for $1 \leqq p \leqq \infty$.

We finally recall from Rothblum and Tan [1985] that for an arbitrary matrix $B \in R^{n \times n}$, vectors $v, a \in R^n$, and norm $\| \ \|$ on R^n,

$$(4.4) \quad \tau(B, \| \ \|, v) \leqq \|B - va^T\|$$

and

$$(4.5) \quad \tau(B, \| \ \|, v) \leqq \|B - va^T\|_2 \leqq \|B - va^T\|_F.$$

Furthermore, the right-hand side of (4.5) is minimized by selecting $a = B^T v / (\|v\|_2)^2$. Thus $\tau(P, \| \ \|, v)$ in (4.3) can be replaced by $\|P - va^T\|$ for any arbitrary vector a in R^n. Furthermore, if $\| \ \|$ is the l_2 norm, $\tau(P, \| \ \|, v)$ can be replaced by $\|P - va^T\|_F$ and this expression is minimized by selecting $a = P^T v / (\|v\|_2)^2$. We further note that when w is the eigenvector of P corresponding to its spectral radius ρ, the spectral radius $\rho(P - wa^T)$ of $P - wa^T$ is known to be an upper bound of $\zeta(P)$, where a is an arbitrary vector satisfying $a^T w \leqq \rho$; again, see Rothblum and Tan [1985]. Thus, using the continuity of the spectrum with the elements of a matrix, one can obtain bounds on $\zeta(P)$ from $\rho(P - va^T)$ when v is an approximation of w. We do not include further details because we were not able to derive simple representations for the tightness of these bounds or use the results to obtain efficient computational methods.

5. Upper bounds on the second largest modulus of an eigenvalue derived from matrices having the same spectrum as the underlying matrix. Given a square, nonnegative, irreducible matrix P, we can apply the methods developed in § 4 to bound the second largest modulus of an eigenvalue of P to any matrix that has the same spectrum as P, e.g., the transpose of P or any scaling of P. Examples of Rothblum and Tan [1985] and Gross [1989] demonstrate that the resulting bounds can be different from those derived from the original matrix (even when the spectral radius and corresponding eigenvector are known and used in the corresponding bounds, rather than their approximations).

We next show that the application of the methods of § 4 to scalings of the given matrix is equivalent to scaling the underlying norm and normalizing vector. In order to describe this fact we need some formal definitions.

Given two matrices P and D in $R^{n\times n}$, where D is diagonal with positive diagonal elements, we define the *D-scaling* of P as the matrix $D^{-1}PD$. Also, if D is as before and $\| \ \|$ is a norm on R^n, we define the *D-scaling* of $\| \ \|$, denoted $\| \ \|_D$, as the norm on R^n with $\|x\|_D = \|Dx\|$ for every $x \in R^n$.

LEMMA 5.1. *Let $D \in R^{n\times n}$ be a diagonal matrix with positive diagonal elements and let $\| \ \|$ be a norm on R^n. Then for every vector $a \in R^n$ and matrix $B \in R^{n\times n}$,*

$$\|a\|_D^* = \|D^{-1}a\|^* = \|a\|_{D^{-1}}^*, \tag{5.1}$$

$$\|D^{-1}BD\| = \|B\|_D, \tag{5.2}$$

and

$$\tau(D^{-1}BD, \| \ \|, a) = \tau(B, \| \ \|_D, Da). \tag{5.3}$$

Proof. Using the formal representation of $\|a\|_D^*$ and a change of variable $x' = Dx$, we have that

$$\|a\|_D^* = \max\{a^Tx : x \in R^n, \|Dx\| \leqq 1\} = \max\{a^TD^{-1}x' : x' \in R^n, \|x'\| \leqq 1\}$$
$$= \|a^TD^{-1}\|^* = \|D^{-1}a\|^* = \|a\|_{D^{-1}}^*,$$

establishing (5.1). Next, by the definition of the functional $\tau(.,.,.)$ and the change of variable $x' = D^{-1}x$,

$$\tau(D^{-1}BD, \| \ \|, a) = \max\{\|x^TD^{-1}BD\| : x \in R^n, \|x\| \leqq 1 \text{ and } x^Ta = 0\}$$
$$= \max\{\|(x')^TBD\| : x' \in R^n, \|Dx'\| \leqq 1 \text{ and } (x')^TDa = 0\}$$
$$= \tau(B, \| \ \|_D, Da),$$

establishing (5.3). The proof of (5.2) follows from the arguments used to establish (5.3) (ignoring the requirement that $x^Ta = 0$). □

Let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding eigenvector w; let $f \in R^n$ satisfy $f^Tw = 1$; and let $\| \ \|$ be a norm on R^n. Also, let D be an $n \times n$ diagonal matrix with positive diagonal elements. Then the spectrum of P and $D^{-1}PD$ coincide; in particular, ρ is the spectral radius of $D^{-1}PD$ with corresponding eigenvector $D^{-1}w$ and $\zeta(P) = \zeta(D^{-1}PD)$. It follows that for a given vector v and a normalizing vector f, bounds on $\zeta(P)$ can be derived by applying Theorem 4.3 to $D^{-1}PD$. We next demonstrate that the resulting bounds are the same bounds derived from the original matrix P with respect to a scaled norm $\| \ \|_D$, where the normalizing vector is $D^{-1}f$ and the approximating vector (for the eigenvector w of P) is Dv.

LEMMA 5.2. *Let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding eigenvector w and let D be an $n \times n$ diagonal matrix with positive diagonal elements. Then ρ is the spectral radius of $D^{-1}PD$ with corresponding eigenvector $D^{-1}w$. Furthermore, let $f \in R^n$ satisfy $f^T(D^{-1}w) = 1$, let $v \in R^n$ satisfy $f^Tv = 1$, and let $\| \ \|$ be a norm on R^n. Then $(D^{-1}f)^T(Dv) = 1$, $\tau(D^{-1}PD, \| \ \|, v) = \tau(P, \| \ \|_D, Dv)$,*

$\|v - D^{-1}w\|^* = \|Dv - w\|_D^*$, $\|Df\| = \|f\|_D$, $\|f^T(D^{-1}PD)\| = \|(D^{-1}f)^TP\|_D$, *and* $\|D^{-1}PD\| = \|P\|_D$.

Proof. The conclusions follow either from the identities of Lemma 5.1 or from the definitions of dual norms and of scalings of norms. □

We next show that applying Theorem 3.1 to a scaling of a given matrix P results in bounds on approximations of the eigenvector of the original matrix P that are equivalently derived from Theorems 3.1, 3.3, and 3.7 applied to P with a corresponding scaling of the given norm.

LEMMA 5.3. *Let P be an $n \times n$, nonnegative, irreducible matrix with spectral radius ρ and corresponding eigenvector w; let D be an $n \times n$ diagonal matrix with positive diagonal elements; and let $\| \ \|$ be a norm on R^n. Also, let f be a positive vector in R^n satisfying $f^T(D^{-1}w) = 1$; let $v \in R^n$ satisfy*

$$D^{-1}PDv \leqq \mu v, \quad D^{-1}PDv \geqq \lambda v, \quad f^Tv = 1, \quad v \geqq 0; \tag{5.4}$$

and let g be the vector defined by $g_i \equiv 1/f_i$, $i = 1, \cdots, n$. Then the bounds on $\|v - D^{-1}w\|$ derived by applying (3.2), (3.7), *and* (3.12) *to $D^{-1}PD$, its spectral radius ρ, and corresponding eigenvector $D^{-1}w$, reduce, respectively, to*

$$\begin{aligned} \|Dv - w\|_{D^{-1}} \leqq (\mu - \rho) \max\ & \|y\|_{D^{-1}} \\ \text{s.t. } & (P - \mu I)y \leqq Dg, \\ & (D^{-1}f)^Ty = 0, \\ & y \in R^n, \end{aligned} \tag{5.5}$$

$$\begin{aligned} \|Dv - w\|_{D^{-1}} \leqq (\rho - \lambda) \max\ & \|y\|_{D^{-1}} \\ \text{s.t. } & (P - \lambda I)y \geqq -Dg, \\ & (D^{-1}f)^Ty = 0, \\ & y \in R^n, \end{aligned} \tag{5.6}$$

and

$$\begin{aligned} \|Dv - w\|_{D^{-1}} \leqq (\mu - \lambda) \max\ & \|y\|_{D^{-1}} \\ \text{s.t. } & (P - \lambda I)y \leqq Dv, \\ & (P - \mu I)y \geqq -Dv, \\ & (D^{-1}f)^Ty = 0, \\ & y \in R^n. \end{aligned} \tag{5.7}$$

Proof. We establish only (5.5), as (5.6) and (5.7) follow analogously. Substituting the explicit expression of the scaled norm $\| \ \|_{D^{-1}}$ into (5.5) and applying the change of variable $y' = D^{-1}y$, we get that (5.5) is equivalent to

$$\begin{aligned} \|v - D^{-1}w\| \leqq (\mu - \rho) \max\ & \|y'\| \\ \text{s.t. } & (D^{-1}PD - \mu I)y' \leqq g, \\ & f^Ty' = 0, \\ & y' \in R^n. \end{aligned} \tag{5.8}$$

The fact that (5.8) is the specialization of (3.2) to $D^{-1}PD$, its spectral radius ρ, and corresponding eigenvector $D^{-1}w$ is obvious; thereby establishing (5.5). □

Acknowledgment. The authors would like to thank Stephen M. Robinson for bringing to their attention the references concerning the computation of the expansion rate of a von Neumann economic model.

REFERENCES

A. BERMAN AND R. J. PLEMMONS [1979], *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York.

L. COLLATZ [1942], *Einschliessungssaetze fuer die charakteristische Zahlen von Matrizen*, Math. Zeitschrift, 48, pp. 221–226.

D. GALE [1960], *The Theory of Linear Economic Models*, McGraw-Hill, New York.

O. GROSS [1989], *Upper bounds on the coefficient of ergodicity*, M.S. thesis, Faculty of Industrial Engineering and Management, Technion, Haifa, Israel.

C. A. HALL AND J. SPANIER [1968], *Nested bounds for the spectral radius*, SIAM J. Numer. Anal., 5, pp. 113–125.

J. M. HAMBURGER, G. L. THOMPSON, AND R. L. WEIL, JR. [1967], *Computation of expansion rates for the generalized von Neumann model of an expanding economy*, Econometrica, 35, pp. 542–547.

T. E. HARRIS [1963], *The Theory of Branching Processes*, Springer-Verlag, Berlin.

M. HAVIV [1988], *Error bounds on approximation to the dominant eigenvector of a nonnegative matrix*, Linear Multilinear Algebra, 23, pp. 159–164.

R. A. HOWARD AND J. E. MATHESON [1972], *Risk sensitive Markov decision processes*, Management Sci., 8, pp. 356–369.

I. MAREK AND R. S. VARGA [1969], *Nested bounds for the spectral radius*, Numer. Math., 14, pp. 49–70.

S. M. ROBINSON [1974], *A linearization technique for solving the irreducible von Neumann economic model*, in Mathematical Models in Economics, J. Los and M. W. Los, eds., Polish Scientific Publisher, Warszawa, pp. 139–150.

U. G. ROTHBLUM [1974], *Multiplicative Markov Decision Chains*, Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, California.

——— [1975], *Algebraic eigenspaces of nonnegative matrices*, Linear Algebra Appl., 12, pp. 281–292.

——— [1981], *Sensitive growth analysis of multiplicative systems* I: *The dynamic approach*, SIAM J. Algebraic Discrete Meth., 2, pp. 25–34.

——— [1984], *Explicit solutions to optimization problems on the intersections of the unit ball of the l_1 and l_∞ norms with a hyperplane*, SIAM J. Algebraic Discrete Meth., 5, pp. 629–632.

U. G. ROTHBLUM AND C. P. TAN [1985], *Upper bounds on the maximum modulus of subdominant eigenvalues of nonnegative matrices*, Linear Algebra Appl., 66, pp. 45–86.

R. S. VARGA [1962], *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

A. F. VEINOTT, JR. [1971], *Class notes in dynamic programming*, Department of Operations Research, Stanford University, Stanford, California.

R. L. WEIL, JR. [1964], *An algorithm for the von Neumann economy*, Zeitschrift fur Nationaloknomie, 24, pp. 371–384.

T. YAMAMOTO [1966], *A computational method for the dominant root of a nonnegative irreducible matrix*, Numer. Math., 8, pp. 324–333.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 33–44, January 1993

003

ON THE SOLUTION TO MATRIX EQUATION $TA - FT = LC$ AND ITS APPLICATIONS*

CHIA-CHI TSUI†

Abstract. Equation $TA - FT = LC$ is fundamental to state space control system design. However, the existing solution to this equation cannot satisfy the necessary requirements that F be in Jordan form and that all rows of T be explicitly expressed in terms of their respective basis vectors. These two requirements are imposed by some basic and practical design problems such as eigenvector assignment and minimal function observer design. This paper shows that a new solution to this matrix equation can truly satisfy those requirements. Consequently, this solution can be applied directly to explicitly solve the above two design problems. Based on this solution, precise loop transfer recovery (LTR), approximate LTR, and output feedback design are unified and much more easily achieved for all systems with more outputs than inputs. This paper also shows the superior efficiency and stability of the computation of this solution.

Key words. state space methods, eigenstructure assignment, minimal function observer design, loop transfer recovery (LTR) observer design, computation

1. Introduction.

Consider the matrix equation

$$TA - FT = LC, \tag{1}$$

where the pair ($A \in R^{n \times n}$, $C \in R^{m \times n}$) is given and is observable, the eigenvalues of $F \in R^{r \times r}$ are arbitrarily given. The problem is to find the solution ($T \in R^{r \times n}$, F, and $L \in R^{r \times m}$) which satisfies (1), where the rows of T must be linearly independent of each other. This equation is fundamental in state space control system design. For example, when $r = n$, (1) is equivalent to

$$T(A - \bar{L}C)T^{-1} = F, \tag{2}$$

which is the state observer design problem with observer gain $\bar{L} = T^{-1}L$. The dual version of (2),

$$T^{-1}(A - BK)T = F, \tag{3}$$

is the state feedback design problem for eigenvalue assignment, where (A, B) is given and is controllable, and K is the state feedback control gain.

It should be noted that when F is in Jordan form, then matrix T will be formed by the left and right eigenvectors (and generalized eigenvectors) of feedback systems (2) and (3), respectively. Thus (1) is equivalent to the entire eigenvalue *and* eigenvector assignment problem.

The existing solution of (1) other than those found in Tsui [16] and Kautsky, Nichols, and Van Dooren [9] can be divided into the following three categories.

(i) In linear systems textbooks, the solution is to first transform (TAT^{-1}, CT^{-1}) into canonical form (Luenberger [12]), and to then adjust L to fit the desired characteristic polynomial coefficients of (2). Therefore, the resulting F of this solution is in companion form (Brand [1]).

(ii) In linear algebraic literature, only the matrix equation

$$TA - FT = C \tag{4}$$

* Received by the editors February 20, 1990; accepted for publication (in revised form) June 5, 1991.

† Department of Applied Sciences, College of Staten Island, City University of New York, Staten Island, New York 10301.

has been treated. Equation (4) is certainly different from (1) in the sense that there is no freedom at the right side of (4) while L of (1) is a freely designed observer gain. Because of this difference in the available freedom, the solution of (4) requires that the eigenvalues of F and A be different, among other restrictions. This is certainly an additional restriction as compared to the arbitrary eigenvalue of F in (1). Therefore, we cannot simply use the solution of (4) to solve (1), even though (4) has already been completely solved, even with good numerical consideration (Golub, Nash, and Van Dooren [8]).

(iii) There are other more complicated algorithms that can be used to solve (1), such as the well-known result of Moore [14], who first showed the freedom of eigenvector assignment.

However, none of these solutions satisfy the following two properties: (i) F is in Jordan form, and (ii) all rows of T are explicitly expressed in terms of their respective basis vectors.

It turns out that the above two properties are *essential* to solving some basic and important design problems, which are more specialized than the simple eigenvalue assignment problem. These problems have not really been solved satisfactorily despite several efforts in the control literature over many years. Three of these problems, the eigenvector assignment for general robustness, the design of reduced-order observers which directly estimate the state feedback, and LTR (recover the loop transfer functions of the observer feedback system), are presented in §§ 4, 5, and 6, respectively. Before these sections on application, a solution of (1) which completely satisfies the above two required properties is presented in §§ 2 and 3, along with its computational analysis.

2. A general solution. The solution that completely satisfies the two required properties of § 1 was independently published in [16] and [9] and is described in this section. The solution of [16], which is more advanced than the solution in this section, is further described in § 3.

The solution is based on the following form of (A, C):

$$(5)\qquad A = [\underset{m}{A_1} : A_2], \quad C = [\underset{m}{C_1} : 0], \quad |C_1| \neq 0,$$

where $|\cdot|$ represents the determinant. Form (5) can be considered general because any C can be transformed to (5) by orthogonal transformation, which does not change the sensitivity of the original problem.

Based on (5), (1) can be partitioned into

$$(6)\qquad (TA - FT)\begin{bmatrix} I_m \\ --- \\ 0 \end{bmatrix} = LC_1$$

and

$$(7)\qquad (TA - FT)\begin{bmatrix} 0 \\ ----- \\ I_{n-m} \end{bmatrix} = 0.$$

For any F and T, (6) can always be satisfied by L because C_1 is full column rank. Thus we only need to solve (7) for F and T. This partitioning shows clearly how the freedom of parameter L can be fully used to reduce the problem dimension from n in (1) to $n - m$ in (7).

Equation (7) can be rewritten as

$$[\mathbf{t}'_1 : \cdots : \mathbf{t}'_r]\left(I_r \otimes A_2 - F' \otimes \begin{bmatrix} 0 \\ \text{-----} \\ I_{n-m} \end{bmatrix} \right) = \mathbf{0}', \tag{8}$$

where $\mathbf{t}'_1$ to $\mathbf{t}'_r$ are the rows of T and $\otimes$ stands for the Kronecker product, and the dimension of this product is $rn \times r(n-m)$. From this dimension, the degree of freedom of solution T of (8) is rm.

With distinctly real eigenvalues, $F = \text{diag}\{\lambda_1, \ldots, \lambda_r\}$. Then from (7) or (8),

$$\mathbf{t}'_i\left(A_2 - \lambda_i \begin{bmatrix} 0 \\ \text{-----} \\ I_{n-m} \end{bmatrix} \right) = \mathbf{0}', \qquad i = 1, \ldots, r, \tag{9}$$

and

$$\mathbf{t}'_i = [c_{i1}, \ldots, c_{im}] \begin{bmatrix} \mathbf{d}'_{i1} \\ \vdots \\ \mathbf{d}'_{im} \end{bmatrix}, \qquad i = 1, \ldots, r, \tag{10}$$

where

$$\mathbf{d}'_{ij}\left(A_2 - \lambda_i \begin{bmatrix} 0 \\ \text{-----} \\ I_{n-m} \end{bmatrix} \right) = \mathbf{0}, \qquad j = 1, \ldots, m. \tag{11}$$

For λ_i and $\lambda_{i+1} = \alpha \pm j\beta$, a corresponding 2×2 diagonal block of the Jordan form F will be formed. Then from (8),

$$[\mathbf{d}'_{ij} : \mathbf{d}'_{i+1,j}]\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes A_2 - \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} \otimes \begin{bmatrix} 0 \\ \text{-----} \\ I_{n-m} \end{bmatrix} \right) = \mathbf{0}', \qquad j = 1, \ldots, m. \tag{12}$$

For $\lambda_i = \lambda_{i+1}$, a corresponding 2×2 diagonal block of F will also be formed. Then from (8),

$$[\mathbf{d}'_{ij} : \mathbf{d}'_{i+1,j}]\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \otimes A_2 - \begin{bmatrix} \lambda_i & 1 \\ 0 & \lambda_i \end{bmatrix} \otimes \begin{bmatrix} 0 \\ \text{-----} \\ I_{n-m} \end{bmatrix} \right) = \mathbf{0}', \qquad j = 1, \ldots, m. \tag{13}$$

Equation (13) implies that

$$\mathbf{d}'_{ij}\left(A_2 - \lambda_i \begin{bmatrix} 0 \\ \text{-----} \\ I_{n-m} \end{bmatrix} \right) = \mathbf{0}' \tag{14}$$

and

$$\mathbf{d}'_{i+1,j}\left(A_2 - \lambda_i \begin{bmatrix} 0 \\ \text{-----} \\ I_{n-m} \end{bmatrix} \right) = \mathbf{d}'_{ij} \begin{bmatrix} 0 \\ \text{-----} \\ I_{n-m} \end{bmatrix}. \tag{15}$$

Equation (15) can be used repeatedly for more repeated λ_i's.

The m $\mathbf{d}_{ij}$ vectors ($j = 1, \ldots, m$) of (11), (12), and (15) always exist because the matrix inside the bracket (and matrix A_2) has full column rank with n rows and $n - m$ columns. These m $\mathbf{d}_{ij}$ vectors form a basis for $\mathbf{t}_i$ as shown in (10).

The above solution completely satisfies the requirements of § 1. Here the arbitrary linear coefficients of (10), c_{ij} ($i = 1, \ldots, r, j = 1, \ldots, m$), form the complete and explicit freedom of T. This result clearly shows why the multivariable system ($m > 1$) has freedom in its eigenvectors T while the single variable system does not.

3. A more advanced solution and its computational analysis. Now, in addition to the form (5), we assume that (A, C) is in the block observable Hessenberg form

$$\left[\begin{array}{c}C\\ \hdashline A\end{array}\right]=\left[\begin{array}{c:c}C_1&0\\ \hdashline A_1&A_2\end{array}\right]=\begin{array}{cc}\left[\begin{array}{c:c:ccc}C_1&0&&&0\\ \hdashline &C_2&0&&0\\ \hdashline &A_{22}&C_3&&\vdots\\ A_1&&&\ddots&\\ &\vdots&&&0\\ &&&&C_v\\ &A_{v2}&&\cdots&A_{vv}\end{array}\right]&\begin{array}{l}m_0\\ m_1\\ m_2\\ \vdots\\ \\ m_{v-1}\\ m_v\end{array}\\ \begin{array}{ccccc}m_1&m_2&m_3&\cdots&m_v\end{array}&\end{array}\tag{16}$$

where $C_1, \ldots, C_v$ are full column rank matrix blocks in echelon form with upper-right corner zeros. Parameter v is the largest observability index, and parameter m_k ($k = 1, \ldots, v$) indicates the number of observability indices that are greater than $k - 1$. Here we have assumed without loss of generality that the observability indices are in descending order. Form (16) can also be considered general because any observable system (A, C) can be transformed to (16) by orthogonal transformation (van Dooren, Emalmi-Naeini, and Silverman [23]).

Because of the echelon form of C_i and therefore A_2, the $\mathbf{d}'_{ij}$ vectors of (11), (12), and (15) can be computed using either back substitution or Givens's method.

The difference between this result and that of § 2 is that observability indices can be used here to analyze the linear dependence of the $\mathbf{d}_{ij}$ vectors. Let us assume that λ_i's are distinct, and let us define the matrix inside the bracket of (11) as $\bar{A}_2$. Then from (16), $\bar{A}_2$ has the same form as A_2. From (16), the jth row of C_{v_j+1} is linearly dependent on its previous rows in C_{v_j+1}, so the corresponding row of $\bar{A}_2$ on its previous rows of $\bar{A}_2$ will be as well. If we express this dependent row of $\bar{A}_2$ as a linear combination of its previous and independent rows of $\bar{A}_2$, then it is straightforward to show that

$$\begin{aligned}\mathbf{d}'_{ij} = [1, \lambda_i, \ldots, \lambda_i^{v-1}]&\left[\begin{array}{c:c:c:c:c}x\cdots x & x\cdots x & \cdots & x\cdots x, *, 0\cdots 0 & 0\cdots 0\\ \hdashline \vdots & \vdots & & & \\ x\cdots x & x\cdots x, *, 0\cdots 0 & & 0 & \\ \hdashline x\cdots x, *, 0\cdots 0 & & & & \\ \hdashline \underbrace{0\ \cdots\ 0}_{m_1} & \underbrace{\qquad\qquad}_{m_2} & \cdots & \underbrace{\qquad\qquad}_{m_{v_j}} & \end{array}\right]\\ &\triangleq \mathbf{v}'_i u_j,\end{aligned}\tag{17}$$

where "x" represents an arbitrary entry and "$*$" represents a nonzero entry and is located at the jth position of the corresponding block.

For example, if $v_1 = 4$, $v_2 = v_3 = 2$, and $v_4 = 1$, then from (16)–(17),

$$A_2 = \left[\begin{array}{ccc:c:c} * & & & & \\ x & * & & & \\ x & x & * & & \\ x & x & x & & \\ \hdashline x & x & x & * & \\ x & x & x & x & \\ x & x & x & x & \\ \hdashline x & x & x & x & * \\ \hdashline x & x & x & x & x \end{array}\right]$$

and

$$\mathbf{d}'_{i1} = \mathbf{v}'_i \begin{bmatrix} x & x & x & x & x & x & x & x & * \\ x & x & x & x & x & x & x & * & 0 \\ x & x & x & x & * & 0 & 0 & 0 & 0 \\ * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{d}'_{i2} = \mathbf{v}'_i \begin{bmatrix} x & x & x & x & x & * & 0 & 0 & 0 \\ x & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{d}'_{i3} = \mathbf{v}'_i \begin{bmatrix} x & x & x & x & x & x & * & 0 & 0 \\ x & x & * & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{d}'_{i4} = \mathbf{v}'_i \begin{bmatrix} x & x & x & * & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where $\mathbf{v}'_i = [1 \; \lambda_i \; \lambda_i^2 \; \lambda_i^3]$.

From (17), it is clear that the m $\mathbf{d}_{ij}$ vectors are linearly independent of each other for a given i ($i = 1, \ldots, r$). Furthermore, for a given j ($j = 1, \ldots, m$), any group of v_j $\mathbf{d}_{ij}$ vectors

$$\begin{bmatrix} \mathbf{d}'_{1j} \\ \vdots \\ \mathbf{d}'_{v_j j} \end{bmatrix} = \begin{bmatrix} \mathbf{v}'_1 \\ \vdots \\ \mathbf{v}'_{v_j} \end{bmatrix} U_j \triangleq V_j U_j$$

are also linearly independent because both V_j and U_j have rank v_j. This property can be easily generalized to the multiple eigenvalue case, where U_j remains unchanged and V_j remains to be a left eigenvector matrix of a companion form matrix (Brand [1]).

An even more special case of the Hessenberg form of (A, C) is the canonical form (Luenberger [12]), for which all "x" entries of A_2 and U_j become 0 and all "$*$" entries become 1. The canonical form cannot be obtained by orthogonal transformation. Nevertheless, the solution based on the canonical form is very suitable for theoretical analysis because it is very simple.

Let us now analyze the computational properties of the algorithm of §§ 2 and 3. The algorithm has only two major steps. The first is the orthogonal transformation of (A, C) into form (5) or (16). This transformation is well known to be efficient and numerically stable, and is a popular first step of many other basic computational problems (such as single value decomposition (SVD) and eigenvalue problems). The second major step is the computation of $\mathbf{d}_{ij}$ vectors in (11), (12), and (15). The computation of (11), (12), and (15) can be direct SVD if based on (5) or direct back substitution if based on (16). This step, especially the back substitution, is well known (Wilkinson [24]) to be direct, simple, numerically stable, and efficient. The efficiency property of this step is further supported by the possibility of completely parallelized computation, which is uniquely enabled by the unique fact that $\mathbf{d}_{ij}$ vectors are completely decoupled for all i and all j if based on (16) (Tsui [20]).

Many control theoreticians and numerical analysts have reservations about the explicit computation of the Jordan form of F and eigenvector matrix T, because computing the Jordan form (the simplest result) from a given matrix can be ill conditioned and very difficult. However, the eigenstructure assignment problem is a *completely different* problem because the eigenvalues are given and need not be computed. It is certainly plausible to use the simplest form (the Jordan form) of these given eigenvalues to compute the rest. In fact, the computation of the eigenvalues is iterative (either QR method or the method of solving characteristic polynomial roots), while the computation of our algorithm is direct and simple.

4. Eigenvector assignment for decoupling and robustness. The first direct application of the solution of §§ 2 and 3 is eigenvector assignment. The only simple *and general* goal of this assignment is to minimize $k(\Lambda) = \|V\|\|V^{-1}\| \geqq 1$, because a small $k(\Lambda)$ generally implies better robustness, smaller control gain, and smoother transient response (Kautsky, Nichols, and Van Dooren [9]), where Λ and V are defined by the equation $V^{-1}(A - BK)V = \Lambda$, with given (A, B) and given eigenvalues of the Jordan form matrix Λ.

To actually achieve (or to try to achieve) $\min_K \{k(\Lambda)\}$ of the above paragraph, we must have explicit solution of V and its complete and explicit freedom (within a variation of orthogonal transformation). In other words, in solving (1) for this problem, F must be within an orthogonal transformation of Λ, and the solution T must be explicit along with complete and explicit freedom. It is obvious that the *only* form of F satisfying this requirement is the Jordan form Λ itself. Other forms of F within the orthogonal transformation of Λ, are too complicated to have the corresponding solution T show complete and explicit freedom.

Indeed, not until 1985 did our solution of §§ 2 and 3, which was based on the Jordan form of F, satisfy the above two requirements. Based on our solution, eigenvector assignment is simplified to the choice of parameters c_{ij} only. In fact, not until 1985 and *only* based on this form of eigenvector assignment freedom, were several explicit numerical algorithms which could compute c_{ij} for a minimum $k(\Lambda)$ developed in Kautsky, Nichols, and Van Dooren [9].

A further improvement on the result of Kautsky, Nichols, and Van Dooren [9] is described in § 3. Because the $\mathbf{d}_{ij}$ vectors in § 2 do not have any analytical property, the computation of their linear combination c_{ij} must be pure numerical iteration in Kautsky, Nichols, and Van Dooren [9]. However, the $\mathbf{d}_{ij}$ vectors in § 3 can be analytically divided into m $(j = 1, \ldots, m)$ output (input in dual case) groups of sizes v_j $(j = 1, \ldots, m)$, which were derived along with the Hessenberg form. As a result, some analytical and direct guidelines of c_{ij} selections for decoupling can be established (Tsui [17]).

For example, if $\lambda_i = \lambda_{i+1}$, we may simply select $\mathbf{t}_i = \mathbf{d}_{ij}$ and $\mathbf{t}_{i+1} = \mathbf{d}_{i,j+1}$, so that the two poles are decoupled into the jth and the $j + 1$th output groups, respectively. By using $\mathbf{d}_{i,j+1}$ instead of $\mathbf{d}_{i+1,j}$ (or the same jth group) for $\mathbf{t}_{i+1}$, a defective eigenvector corresponding to λ_{i+1} and (15) is avoided (Golub and Wilkinson [7]). As another example, if λ_n is closest to the imaginary axis (or dominant), we may let $\mathbf{t}_n = \mathbf{d}_{nj}$ with the smallest v_j. The resulting $\mathbf{t}_n$ will be directly affected by the least number of other eigenvectors. Let us define the sensitivity (or the condition number $k(\lambda_i)$) of an *individual* eigenvalue λ_i as $k^{-1}(\lambda_i) = \langle \bar{\mathbf{t}}_i, \bar{\boldsymbol{t}}_i \rangle \leqq 1$ $(i = 1, \ldots, n)$, where $\bar{\mathbf{t}}_i$ and $\bar{\boldsymbol{t}}_i$ are the ith normalized vectors of T and T^{-1}, respectively [24]. For example, if λ_n is decoupled into a group with the smallest possible size $v_j = 1$, then a 1×1 diagonal block will be created for λ_n, and the corresponding $k(\lambda_n)$ will be a perfect 1. Therefore, our decoupling can uniquely reduce $k(\lambda_i)$.

To guide the above selection of c_{ij} and decoupling, we will establish a new robustness measure. From [9], robustness usually implies small control gain and smooth transient response. Robustness is the sensitivity of the system toward instability. The often used general robustness measure, which involves pole sensitivity, is $S_1 = k^{-1}(\Lambda)|\text{Re}(\lambda_n)|$. However, $k(\Lambda)$ is only an *overall* sensitivity of *all* eigenvalues, while the sensitivities of some individual λ's (such as the dominant ones) should be more important in robustness than the other λ's, as shown by the example of Lewkowicz and Sivan [11]. Thus our new robustness measure is proposed as

$$S_2 = \min_i \{k^{-1}(\lambda_i)|\text{Re}(\lambda_i)|\}. \tag{18}$$

Because $k(\lambda_i) = \|\mathbf{t}_i\|\|\underline{\mathbf{t}}_i\| < \|T\|\|T^{-1}\| = k(\Lambda)$, where $\mathbf{t}_i$ and $\underline{\mathbf{t}}_i$ are the ith vector of T and T^{-1}, respectively, and because $k(\lambda_i) \geqq 1$ for all i, it is clear that

$$k^{-1}(\Lambda)|\text{Re}(\lambda_n)| < S_2 \leqq |\text{Re}(\lambda_n)|. \tag{19}$$

Comparing S_1 with S_2, we see that instead of using the overall pole sensitivity $k(\Lambda)$ in S_1, S_2 uses the sensitivity of λ_n *itself* to divide $|\text{Re}(\lambda_n)|$, as the likely distance for λ_n to become unstable. Therefore, S_2 is always more accurate and less conservative than S_1. Furthermore, unlike S_1, S_2 considers the sensitivities and real parts of *all* individual λ's, not just that of λ_n. It should be noted that the instability of *any* λ will cause the instability of the system. Therefore, S_2 considers more complete information than S_1. Nevertheless, S_2 is still less conservative than S_1, as shown by the lower bound of (19). Finally, S_2 can be easily modified for different applications. S_2 considers the weighted sensitivities of each individual pole. The heaviest weighting factor $|\text{Re}(\lambda_n)|$ is on maximizing $k^{-1}(\lambda_n)$, so that the product pair will not be the minimum among the n pairs in S_2 of (18). We may arbitrarily modify these weighting factors if we know a priori that the sensitivities of some individual poles are particularly important (or less important).

The unique feature of S_2 is its use of individual pole sensitivities. The decoupling can uniquely reduce these sensitivities (see [11]). The guidelines of our eigenvector assignment can uniquely and directly achieve decoupling and, therefore, the enhanced robustness (in the direction of S_2). These guidelines, like pole selections, are used to guide designer/computer interaction and iteration, but not a one-call computer subroutine. Because of the variety and complexity of the real-world systems, as indicated by different v_j's, m's, and p's, and because of the complicated trade-off between the desired performance (indicated by λ's) and robustness (indicated by λ's, $k(\lambda_i)$'s, and S_2), it is necessary to compute the real optimal solution interactively. Our result is aimed directly at improving this designer/computer interaction, but not at replacing this interaction (by a one-call computer subroutine).

5. Simplifying minimal function observer design to the solving of a set of linear equations. The second application of our solution of (1) is the design of a general observer

$$\begin{aligned} \dot{\mathbf{z}}(t) &= F\mathbf{z}(t) + L\mathbf{y}(t) + TB\mathbf{u}(t), \\ K\mathbf{x}(t) &= N\mathbf{z}(t) + M\mathbf{y}(t), \end{aligned} \tag{20}$$

where $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{u}$ are the system states, outputs, and inputs, respectively. Equation (20) is a general formulation of an observer which estimates $K\mathbf{x}(t)$. A state observer is a special case of (20) when $K = I$. The observer (20) is also defined as "Kalman type" if $M = 0$ and "Luenberger type" if $M \neq 0$. For example, the Kalman filter, with (K, M) fixed to be $(I, 0)$, is a special case of (20).

The *real* original purpose of an observer is to estimate $K\mathbf{x}(t)$, *not* $\mathbf{x}(t)$. When $\mathbf{x}(t)$ is estimated by some observers, it will be immediately multiplied by K anyway. Because the number of $K\mathbf{x}(t)$ (the number of control inputs) is usually much less than the number of states $\mathbf{x}(t)$, directly estimating $K\mathbf{x}(t)$ should require much lower observer order than estimating $\mathbf{x}(t)$. A general observer (20) that estimates $K\mathbf{x}(t)$ directly is called a function observer.

Equation (1), which implies that $\mathbf{z}(t) \Rightarrow T\mathbf{x}(t)$, is a necessary condition for all types of observers. Substituting $\mathbf{z}(t) = T\mathbf{x}(t)$ into (20), the sufficient condition is then

$$K = NT + MC = [N : M] \begin{bmatrix} T \\ \text{---} \\ C \end{bmatrix} \begin{matrix} r \\ \\ m \end{matrix} . \tag{21}$$

Although equations (1) and (21) are both the necessary and sufficient conditions of observer (20) [13], (21) (not (1)) exclusively reflects the difference (in K and M) of all types of observers. Therefore, (21) should be used exclusively to systematically determine the different orders of different types of observers. Therefore, the minimum observer order r is determined as the minimum r needed to solve (21) for N and M.

For example, in state observers ($K = I$), the observer order r is n for $M = 0$ and $n - m$ for $M \neq 0$, and is independent of the values of T and C. However, in function observers, because K usually has much smaller numbers of rows (state feedbacks) than I, the corresponding r can be much smaller. Furthermore, in function observers, the minimum r is dependent on all values of K, T, and C. Thus the minimum r must be determined from (21) for each case of K, T, and C. Because K is arbitrarily given and T is determined from (1) with n^2 entries, the explicit and analytical formula for minimum r is a function of n^2 variables and therefore does not exist.

To actually determine the minimum r from (21) systematically, the eigenvalues of F and the corresponding rows of T must be completely decoupled. Otherwise, reducing one observer order or extracting one eigenvalue from F (or one row from T of (21)) will change the remaining eigenvalues of F and T. Consequently, before 1985, because of the difficulty of deriving such an F and T, people were forced to solve (1) and (21) together. Only based on the solution of (1) in §§ 2 and 3, can we now solve (21) exclusively and systematically, for a minimum r [16].

Because (21) alone is only a set of linear equations, it is itself the simplest possible theoretical solution of minimal observer design. The actual solving of this set of linear equations is technical. Although the minimum r does not have a general and explicit formula because it depends on too many variables, its upper bound has been derived by solving (21) exclusively [18]. The lower bound of r is 0, also from (21).

The claim made in the title of this section has never been made before and has been unnoticed by people studying this basic problem (see, e.g., Fowell, Bender, and Assal

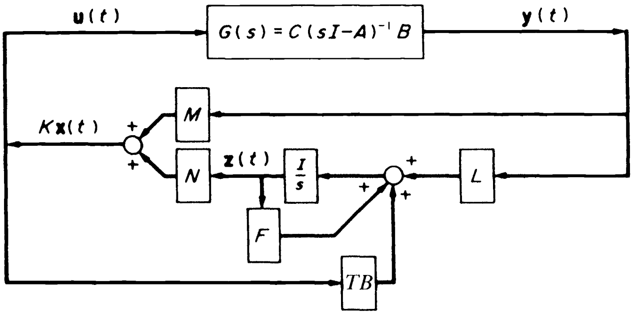


FIG. 1

[6]). This is probably because the result in [16] and [18] is not as simple as the result of state observers, despite the fact that the function observer is much more complicated (r depends on K, T, and C) and the fact that (21) is already the simplest possible theoretical solution.

6. Unified precise LTR and output feedback design. The third application of our solution of (1) is loop transfer recovery (LTR). LTR is an additional design requirement of observer (20), whose corresponding block diagram is shown in Fig. 1. LTR requires the observer (20) not only to estimate $K\mathbf{x}(t)$, but also to have the corresponding loop transfer function $L_0(s)$ (at the break point $\mathbf{u}(t)$) be recovered to the loop transfer function of the direct state feedback system $L_s(s) = K(sI - A)^{-1}B$. It was pointed out by Doyle [3] that $L_0(s)$, unlike the loop transfer function $L(s)$ at the break point $K\mathbf{x}(t)$, is different from $L_s(s)$.

Because the properties of a feedback control system (such as sensitivity and robustness) are contained by the system's loop transfer function, LTR actually requires the observer not only to implement the given state feedback control, but also to preserve the properties of the corresponding state feedback system. Therefore, LTR is a very practical and important problem.

The sufficient condition of LTR based on observer formulation (20) is proved to be $TB = 0$ [19], [21]. This condition can be directly visualized from the above block diagram, which shows that the difference between the break points $\mathbf{u}(t)$ and $K\mathbf{x}(t)$ is caused by the feedback path TB.

The existing LTR approach (Doyle and Stein [4], [5]) uses Kalman filter as the observer. While the complete freedom of observer design can be represented by observer eigenstructure assignment, the only explicitly used freedom in Kalman filter design (for LTR) is the increase of system input noise level. Based on this approach, precise LTR ($L_0(s) = L_s(s)$) with finite filter gain can be achieved only when the open loop system $G(s)$ has $n - m$ stable invariant zeros, and only if all $n - m$ Kalman filter poles match each one of these zeros. For all systems not having $n - m$ stable invariant zeros, only approximate LTR can be achieved and *only* with asymptotically large filter gain. This result has received wide and continuous attention (see, e.g., Stein and Athans [15]).

There is another result about precise LTR ($TB = 0$), which is referred to as "unknown input observers." Let us quote the latest treatment of that result (Kudva, Viswanadham, and Ramakrishna [10]), which says that for all minimum-phase systems with $m > p$ and rank $(CB) = p$, precise LTR ($TB = 0$) with finite observer gain can be achieved. In addition, if the above system does not have any invariant zeros, then the observer poles can be arbitrarily chosen. No approximate result has been derived for systems not satisfying the above conditions, and this result is completely different from that of Doyle and Stein [4], [5].

From our analysis of LTR at the beginning of this section, the LTR problem can be formulated to satisfy (1), (21), and $TB = 0$. Now based on our solution of (1) in §§ 2 and 3, we only need to satisfy (21) and $TB = 0$ by fully using the remaining observer design freedom of eigenvalue selection and eigenvector (c_{ij}) assignment. Now, because for $i = 1, \dots, n-m$,

$$(22) \qquad \mathbf{t}'_i B = [c_{i1} \cdots c_{im}] \begin{bmatrix} \mathbf{d}'_{i1} \\ \vdots \\ \mathbf{d}'_{im} \end{bmatrix} B \triangleq [c_{i1} \cdots c_{im}](D_i B)_{m \times p'},$$

we can always make (22) = 0 by selecting the free parameters c_{ij} if $m > p$ and if the observer poles are distinct and real. We will further assign the observer poles exactly the same way as in [10], that is, we will match every existing invariant zero of $G(s)$ (they are stable, as guaranteed by minimum phase). Thus we have followed all design requirements of [10] and fully used the entire remaining observer design freedom (eigenvector (c_{ij}) assignment). If $G(s)$ has $n - m$ stable invariant zeros, then $D_i B = N(s = \lambda_i) = 0$ for all i and for *any* values of (A, B, C), and m and p, where $G(s)$ = left MFD = $D^{-1}(s)N(s)$ (see Chen [2]). In this case, we can arbitrarily select c_{ij} to make the corresponding matrix $[T' : C']$ full rank or to make (21) solvable for any given K, and we can always find such c_{ij} directly (Tsui [22]). Therefore, we have easily and uniquely achieved the precise LTR results of *both* Doyle and Stein [4], [5] and Kudva, Viswanadham, and Ramakrishna [10].

For systems not satisfying the conditions in [4], [5], and [10], precise LTR with finite observer gain cannot be achieved. As reflected in our design for satisfying (1), (21), and $TB = 0$, (21) and $TB = 0$ may not both be satisfied for those systems, while (1) is always satisfied, based on the solution of §§ 2 and 3. Now, for any systems with $m > p$ and distinct and real observer poles, $TB = 0$ can always be satisfied, as shown in (22). For those systems, only (21) is not guaranteed solvable or $[T' : C']$ is singular. In this case, we can simply find an approximate solution of (21), $\bar{K} = [N : M][T' : C']'$, with T determined from (22) and $\bar{K}$ only approximating K (as in the least-square solution of (21)). The resulting loop transfer function will be $\bar{K}(sI - A)^{-1}B$, and only $\bar{K}\mathbf{x}(t)$ (not the desired $K\mathbf{x}(t)$) will be estimated and implemented. Therefore, we achieved approximate LTR more systematically and with guaranteed finite observer gain, because (21) and $TB = 0$ are only sets of linear equations.

There is a very important modification or extension of our LTR procedure, also valid for all systems with $m > p$ and distinct and real observer poles. Because our solution T of (1) and (22) ($TB = 0$) is uniquely *independent* of K, we can design a *new* $\bar{K}\mathbf{x}(t)$ instead of estimating or approximating a given $K\mathbf{x}(t)$, based on the available $\bar{\mathbf{y}}(t) \triangleq [\mathbf{z}(t)' : \mathbf{y}(t)']' = [T' : C']'\mathbf{x}(t)$ in (20).

It turns out that this modification is equivalent to output feedback (since $TB = 0$) design, which has been extensively studied recently and has shown promising results. Here we are designing an output feedback gain $[N : M]$ from the given $\mathbf{y}(t) = C\mathbf{x}(t)$ *and* $\bar{\mathbf{z}}(t) = \bar{T}\mathbf{x}(t)$, where $\bar{T}$ is a part of T which makes $[\bar{T}' : C']'$ full row rank and $\bar{\mathbf{z}}(t)$ is independent of $\mathbf{y}(t)$ [2]. Because of the unique decoupled property of all rows of T of our solution, $\bar{T}$ can be directly extracted from T. We call this modified LTR design an improved output feedback design because we are using an "additional output" $\bar{\mathbf{z}}(t)$ instead of using $\mathbf{y}(t)$ only. Because $\bar{\mathbf{z}}(t)$ constitutes the dynamic part of output feedback

($TB = 0$) observer of (20), we can also consider this improved output feedback as dynamic output feedback.

When a system satisfies the conditions of either Doyle and Stein [4], [5] or Kudva, Viswanadham, and Ramakrishna [10], $[T' : C']$ will be full rank and *any* state feedback $Kx(t)$ can be estimated and implemented (with $TB = 0$). Thus, uniquely based on the solution of (1) in §§ 2 and 3, LTR is directly modified as an improved output feedback problem, and is perfectly unified with the output feedback design. This design and estimation centered on (21) also nicely unifies with the reduced-order observer design of § 5.

7. Conclusion. Both the Lyapunov equation ($TA - A'T = C$) and Sylvester equation ($TA - FT = C$) are classical matrix equations. It is surprising that equation $TA - FT = LC$ is much less understood, despite its fundamental importance in state space design theory.

In this paper, we have shown the technical and computational advantages of our new solution to the equation $TA - FT = LC$. In addition, this solution is extremely simple, especially in light of its capability for explicitly solving eigenvector assignment, function observer design, and LTR design problems. These three design problems are very basic and practical. Therefore, we hope this result finds its way into future textbooks and computer-aided design packages about controller design using state space techniques.

Acknowledgment. The author wishes to thank the reviewers of this paper for providing many suggestions for improving its presentation.

REFERENCES

[1] L. Brand, *The companion matrix and its properties*, Amer. Math. Monthly, 75 (1968), pp. 146–152.

[2] C. T. Chen, *Linear System Analysis and Design*, Holt, Rinehart, and Winston, New York, 1984.

[3] J. Doyle, *Guaranteed stability margins for* LQG *regulators*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 756–757.

[4] J. Doyle and G. Stein, *Robustness with observers*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 607–611.

[5] J. Doyle, *Multivariable feedback design concepts for classical/modern analysis*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 4–16.

[6] R. Powell, D. Bender, and A. Assal, *Estimating the plant state from the compensator state*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 964–967.

[7] G. Golub and H. Wilkinson, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev., 18 (1976), p. 578.

[8] G. Golub, S. Nash, and P. van Dooren, *A Hessenberg–Schur method for the problem $AX - SB = C$*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 909–912.

[9] J. Kautsky, N. K. Nichols, and P. van Dooren, *Robust pole assignment in linear state feedback*, Internat. J. Control, 41 (1985), pp. 1129–1155.

[10] P. Kudva, N. Viswanadham, and A. Ramakrishna, *Observers for linear systems with unknown inputs*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 113–115.

[11] I. Lewkowicz and R. Sivan, *Maximal stability robustness for state equations*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 297–300.

[12] D. G. Luenberger, *Canonical forms for linear multivariable systems*, IEEE Trans. Automat. Control, AC-16 (1967), pp. 290–293.

[13] ———, *Introduction to observers*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 596–603.

[14] B. C. Moore, *On the flexibility offered by state feedback in multivariate systems beyond closed loop eigenvalue assignment*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 689–692.

[15] G. Stein and M. Athans, *The* LQG/LTR *procedure for multivariable feedback control design*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 105–114.

[16] C. C. TSUI, *A new algorithm for the design of multifunctional observers*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 89–93.

[17] ———, *An algorithm for computing state feedback in multi-input linear systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 243–246.

[18] ———, *On the order reduction of linear function observers*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 447–449.

[19] ———, *On preserving the robustness of optimal control system with observers*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 823–826.

[20] ———, *A new approach to robust observer design*, Internat. J. Control, 47 (1988), pp. 745–751.

[21] ———, *On robust observer compensator design*, Automatica, 24 (1988), pp. 687–692.

[22] ———, *Recent advances in control*, IEEE Press Book, 1990, pp. 193–195.

[23] P. VAN DOOREN, A. EMAIMI-NAEINI, AND L. SILVERMAN, *Stable extraction of the Kronecker structure of pencils*, in Proc. 17th IEEE Conf. Decision and Control, 1978, pp. 521–524.

[24] H. WILKINSON, *Algebraic Eigenvalue Problem*, Oxford University Press, London, U.K., 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 45–58, January 1993

004

FACTORIZED SPARSE APPROXIMATE INVERSE PRECONDITIONINGS I. THEORY*

L. YU. KOLOTILINA† AND A. YU. YEREMIN‡

Abstract. This paper considers construction and properties of factorized sparse approximate inverse preconditionings well suited for implementation on modern parallel computers. In the symmetric case such preconditionings have the form $A \rightarrow G_L A G_L^T$, where G_L is a sparse approximation based on minimizing the Frobenius form $\|I - G_L L_A\|_F$ to the inverse of the lower triangular Cholesky factor L_A of A, which is not assumed to be known explicitly. These preconditionings preserve symmetry and/or positive definiteness of the original matrix and, in the case of M-, H-, or block H-matrices, lead to convergent splittings.

Key words. sparse approximate inverse, preconditioning, M-, H-, block H-matrices, convergent splittings

AMS(MOS) subject classifications. 65F10, 65F35, 65F50, 65Y05

1. Introduction.

This paper is the first in a series of papers (see also [10] and [11]) on the same topic. The papers are dedicated to the construction of a special type of explicit preconditioning based on factorized sparse approximate inverses and to their application for solving three-dimensional finite element systems on massively parallel computers.

In the context of modern massively parallel computers, it is important to consider explicit preconditionings of linear algebraic systems. Such preconditionings, requiring additionally only matrix–vector multiplications at each iteration of the conjugate gradient (CG) method, or another basic iterative method, and thus possessing a natural parallelism, were suggested in [2] and extended in [6]–[9]. The main idea exploited in these papers is as follows. Let a system

$$Ax = b \tag{1.1}$$

with a positive definite (not necessarily symmetric) $n \times n$ matrix A be given. To accelerate convergence of a basic iterative method we use explicit preconditioning, i.e., we pass from (1.1) to an equivalent system

$$GAx = Gb, \tag{1.2}$$

where G is a nonsingular sparse approximate inverse of A.

To construct $G = (g_{ij})$, we prescribe its sparsity pattern $S \subseteq \{(i,j) : 1 \leqq i \neq j \leqq n\}$, i.e., we set

$$g_{ij} = 0 \quad \text{if } (i,j) \in S, \tag{1.3}$$

while for $(i, j) \notin S$, g_{ij} are considered free parameters. To determine their values, we choose a positive definite, possibly nonsymmetric weight matrix W and minimize the nonnegative quadratic functional

$$\|I - GA\|_W^2 = \operatorname{tr}\,[(I - GA)W(I - GA)^T]. \tag{1.4}$$

* Received by the editors June 16, 1990; accepted for publication (in revised form) June 16, 1991.

† V. A. Steklov Mathematical Institute, Russian Academy of Sciences, Fontanka 27, St. Petersburg 191011, Russia (liko@lomi.spb.su).

‡ Department of Numerical Mathematics, Russian Academy of Sciences, Leninskij Prospect, 32A, 117334 Moscow, Russia (badger@adonis.ias.msk.su).

Recall that a nonsymmetric matrix A is said to be positive definite if its symmetric part $\frac{1}{2}(A + A^T)$ is positive definite. This leads to the following equations for nonzero entries of G:

$$(GA(W + W^T)A^T)_{ij} = ((W + W^T)A^T)_{ij}, \qquad (i, j) \notin S. \tag{1.5}$$

Clearly, (1.5) decouples into n independent linear systems with symmetric positive definite coefficient matrices, each of which is used to compute nonzero entries in one row of G.

When A is symmetric positive definite, the choice $W = A^{-1}$ is possible and (1.5) takes the form

$$(GA)_{ij} = \delta_{ij}, \qquad (i, j) \notin S, \tag{1.6}$$

where δ_{ij} is the Kronecker symbol. Obviously, (1.6) is consistent whenever each principal submatrix of A is nonsingular. This enables us to utilize (1.6) for constructing sparse approximate inverses, e.g., for any (not necessarily symmetric) positive definite matrix A. As shown in [7], G is also well defined by (1.6) when A is an H-matrix. Furthermore, in this case, the preconditioned matrix GA is again an H-matrix and the related splitting

$$A = G^{-1} - R$$

is convergent, i.e., $\rho(I - GA) < 1$.

Unfortunately, (1.6) (as well as (1.5) for $W \neq A^{-2}$) generally yields an unsymmetric approximate inverse G even if the original matrix A and the sparsity pattern S are symmetric. Thus a real spectrum of A might be transformed under such preconditioning into a complex spectrum of GA. Some methods for constructing symmetric approximate inverses were presented in [8].

This paper provides the theoretical background for factorized sparse approximate inverse preconditionings based on minimizing the Frobenius matrix norm and preserving symmetry and/or positive definiteness of the original matrix. In the symmetric case such preconditionings appeared in [5].

In [10] we specify a large class of finite element applications (related to the three-dimensional elasticity problems), and consider efficient implementation of sparse approximate inverse-preconditioned CG iterations on a model parametrized, massively parallel, hypercube computer. The numerical results presented therein show that under an appropriate selection of the preconditioner sparsity pattern, we can ensure a high convergence rate of the preconditioned CG iterations (comparable with that of the best serial preconditioning methods). Furthermore, different strategies for selecting preconditioner sparsity patterns for three-dimensional finite element systems are compared.

Further methods of the sparsity pattern selection for three-dimensional finite element problems aimed at accelerating convergence of the CG iterations are suggested in [11]. Because direct construction of such preconditioners is rather computationally expensive, the paper suggests applying preconditioned block CG iterations for computing block rows of the former. The numerical results presented therein show that a few preconditioned block CG iterations are required for constructing approximate preconditioners whose preconditioning quality only slightly differs from that of the corresponding exact preconditioners.

This paper is organized as follows. The general construction of factorized sparse approximate inverse preconditionings is described in § 2. Section 3 presents an approximation property of staircase, structurally symmetric, factorized sparse approximate inverses. In § 4 we prove that factorized sparse approximate inverse preconditionings preserve the property of a matrix to be an M-, H-, or block H-matrix, and establish convergence results.

We conclude this introduction by specifying some notation. For a real $m \times n$ matrix $A = (a_{ij})(m, n \geqq 1)$, the inequality $A > 0$ ($A \geqq 0$) is componentwise, i.e., $a_{ij} > 0$ ($a_{ij} \geqq 0$) for all i and j, $1 \leqq i \leqq m$, $1 \leqq j \leqq n$. By $|A|$ we denote the matrix whose entries are absolute values of the corresponding entries of A, i.e., $|A|_{ij} = |a_{ij}|$. For an $n \times n$ matrix $A = (a_{ij})$ by $\operatorname{diag}(A) = \operatorname{diag}(a_{11}, \ldots, a_{nn})$, we denote the diagonal matrix whose diagonal entries coincide with those of A. The block diagonal matrix block diag (A) = block diag $(A_{11}, \ldots, A_{nn})$ for a block $n \times n$ matrix $A = (A_{ij})$ is defined analogously. Finally, by tr A and $\rho(A)$ we denote, respectively, the trace and the spectral radius of a square matrix A.

Note also that by stating that a matrix A is an M-matrix, we assume that A is nonsingular.

2. Construction of factorized sparse approximate inverses. Let A by a symmetric, positive definite matrix and

$$A = L_A L_A^T \tag{2.1}$$

be its Cholesky factorization. We look for an approximate inverse of A in the factorized form

$$H = G_L^T G_L, \tag{2.2}$$

where G_L is a sparse nonsingular lower triangular matrix approximating L_A^{-1}. Such a form of an approximate inverse is related to the preconditioning

$$A \rightarrow G_L A G_L^T,$$

clearly preserving symmetry and/or positive definiteness of A.

To construct G_L, we first fix a sparsity pattern $S_L \subseteq \{(i, j) : 1 \leqq i \neq j \leqq n\}$ such that $S_L \supseteq \{(i, j) : i < j\}$ and determine nonzero entries of G_L (located in the lower triangular part) using (1.5) with A replaced by L_A and W replaced by I, i.e., by minimizing the Frobenius norm $\| I - G_L L_A \|_F$. Taking into account (2.1), (1.5) can be then rewritten as

$$(G_L A)_{ij} = (L_A^T)_{ij}, \qquad (i, j) \notin S_L. \tag{2.3}$$

Note that by the choice of S_L, the condition $(i, j) \notin S_L$ implies that $i \geqq j$, while L^T is upper triangular. We can thus present (2.3) in the equivalent form

$$(G_L A)_{ij} = \begin{cases} 0, & i \neq j, \quad (i, j) \notin S_L, \\ (L_A)_{ij}, & i = j. \end{cases} \tag{2.4}$$

Relations (2.4) show that up to a left diagonal scaling factor, the matrix G_L approximating L_A^{-1} coincides with the (lower triangular) sparse approximate inverse of A of type (1.6) and off-diagonal entries of L_A are luckily not required to be known at all.

If the diagonal entries of L_A are not available, we construct G_L as follows. First compute the matrix $\hat{G}_L$ of the same sparsity pattern as G_L using the equations

$$(\hat{G}_L A)_{ij} = \delta_{ij}, \qquad (i, j) \notin S_L, \tag{2.5}$$

and set $G_L := D\hat{G}_L$, where the diagonal scaling matrix D is so chosen that

$$(G_L A G_L^T)_{ii} = 1. \tag{2.6}$$

In other words, the matrix $\hat{G}_L A \hat{G}_L^T$ is symmetrically Jacobi scaled to produce the ultimate preconditioned matrix.

It is, of course, possible to apply another strategy for scaling $\hat{G}_L A \hat{G}_L^T$, e.g., we can select D to minimize $\|I - G_L A G_L^T\|_F$. Our choice is motivated by the following well-known results.

THEOREM 2.1 [15]. *Let A be a symmetric positive definite matrix and $D = \text{diag}\,(A)$. Then*

$$\text{COND}\,(D^{-1/2} A D^{-1/2}) \leqq m\,\text{COND}\,(\tilde{D} A \tilde{D}),$$

where m is the maximum number of nonzeros in any row of A, $\tilde{D}$ is an arbitrary positive definite diagonal matrix, and COND $(B) = \lambda_{\max}(B)/\lambda_{\min}(B)$ *is the spectral condition number of B.*

THEOREM 2.2 [4]. *If, under the hypotheses of Theorem* 2.1, *the matrix A has property $\mathscr{A}$, then*

$$\text{COND}\,(D^{-1/2} A D^{-1/2}) \leqq \text{COND}\,(\tilde{D} A \tilde{D}).$$

Furthermore, as we shall see below, G_L, constructed using (2.5) and (2.6), minimizes the functional

$$\text{tr}\,(XAX^T)/\text{det}^{1/n}\,(XAX^T) \tag{2.7}$$

over all matrices X of the same sparsity pattern S_L. This functional was considered in [5], and it can be viewed as a measure of closeness of XAX^T to the identity matrix.

Let us consider relations (2.5) and (2.6) in more detail. Define, for a sparsity pattern S, the matrices $A^{(i)}(S)$, $i = 1, \ldots, n$, in the following way:

$$(A^{(i)}(S))_{rs} = \begin{cases} a_{rr}, & r = s, \\ a_{rs}, & (i, r) \notin S \quad \text{and} \quad (i, s) \notin S, \\ 0, & \text{otherwise}. \end{cases} \tag{2.8}$$

Then, obviously, the ith row of $\hat{G}_L$ coincides with the ith row of $[A^{(i)}(S_L)]^{-1}$. Furthermore, since

$$(\hat{G}_L A \hat{G}_L^T)_{ii} = \sum_{j:(i,j)\notin S_L} \delta_{ij} (\hat{G}_L)_{ij} = (\hat{G}_L)_{ii} = ([A^{(i)}(S_L)]^{-1})_{ii},$$

relation (2.6) implies that

$$D = [\text{diag}\,(\hat{G}_L)]^{-1/2}, \tag{2.9}$$

and thus relations (2.5) and (2.6) can be summarized in the form

$$(G_L A)_{ij} = \delta_{ij}/([A^{(i)}(S_L)]^{-1})_{ii}^{1/2}, \qquad (i, j) \notin S_L, \tag{2.10}$$

which coincides with the formula in [5] derived by minimizing functional (2.7).

The above construction of factorized sparse approximate inverse preconditionings can be readily extended to the unsymmetric case. Assume that each principal submatrix of A is nonsingular (which is the case, for example, if A is positive definite, i.e., $A + A^T$ is symmetric positive definite).

Select some triangular sparsity patterns

$$\{(i, j) : i < j\} \subseteq S_L \subseteq \{(i, j) : i \neq j\} \tag{2.11a}$$

and

$$\{(i, j) : i > j\} \subseteq S_U \subseteq \{(i, j) : i \neq j\}. \tag{2.11b}$$

The lower and upper triangular matrices G_L and G_U are constructed to satisfy the conditions

$$(2.12)\qquad \begin{aligned} (G_L)_{ij} &= 0, & (i,j) &\in S_L, \\ (G_U)_{ij} &= 0, & (i,j) &\in S_U, \\ (G_LA)_{ij} &= \delta_{ij}, & (i,j) &\notin S_L, \\ (AG_U)_{ij} &= \delta_{ij}, & (i,j) &\notin S_U, \end{aligned}$$

and the preconditioned matrix is presented in the form

$$(2.13)\qquad G_LAG_UD^{-1},$$

where

$$(2.14)\qquad D = \operatorname{diag}(G_LAG_U).$$

Note that since

$$(G_LAG_U)_{ii} = \sum_{k \leq i} (G_LA)_{ik}(G_U)_{ki} = (G_U)_{ii} + \sum_{\substack{k<i \\ (i,k)\in S_L \\ (k,i)\notin S_U}} (G_LA)_{ik}(G_U)_{ki},$$

we have

$$(2.15a)\qquad (G_LAG_U)_{ii} = (G_U)_{ii}$$

if there exist no $k < i$ such that $(i, k) \in S_L$ and $(k, i) \notin S_U$. Similarly, the equality

$$(G_LAG_U)_{ii} = (G_L)_{ii} + \sum_{\substack{k<i \\ (i,k)\notin S_L \\ (k,i)\in S_U}} (G_L)_{ik}(AG_U)_{ki}$$

yields

$$(2.15b)\qquad (G_LAG_U)_{ii} = (G_L)_{ii}$$

if there exist no $k < i$ such that $(i, k) \notin S_L$ and $(k, i) \in S_U$.

On the other hand, since the ith diagonal entries $(G_L)_{ii} = ([A^{(i)}(S_L)]^{-1})_{ii}$ and $(G_U)_{ii} = ([A^{(i)}(S_U)]^{-1})_{ii}$ coincide with the corresponding diagonal entries of the matrices which are positive definite if A is positive definite, the diagonal entries of D are positive whenever $(G_LAG_U)_{ii} = (G_L)_{ii}$ or $(G_LAG_U)_{ii} = (G_U)_{ii}$ for every $i = 1, \ldots, n$. We have thus proved the following.

LEMMA 2.1. *Let A be positive definite and the sparsity patterns S_L and S_U satisfy* (2.11) *and the following condition: for every i, $i = 1, \ldots, n$, either there exist no $k < i$ such that $(i, k) \in S_L$ and $(k, i) \notin S_U$, or there exist no $k < i$ such that $(i, k) \notin S_L$ and $(k, i) \in S_U$. Then the matrices G_L, G_U, and D defined by* (2.12) *and* (2.14) *have positive diagonal entries and thus preconditioning* (2.13) *is well defined.*

Noting that the hypotheses on S_L and S_U of Lemma 2.1 are trivially satisified when $S_U = S_L^T$, and taking into account (2.15), we obtain the following corollary.

COROLLARY 2.1. *If S_L satisfies* (2.11) *and $S_U = S_L^T$, then*

$$D = \operatorname{diag}(G_L) = \operatorname{diag}(G_U).$$

3. An approximation property of factorized sparse approximate inverses. In this section we establish an interesting property of the factorized sparse approximate inverse $G_UD^{-1}G_L$ of a positive definite matrix A.

Let us introduce some notation. For a lower triangular sparsity pattern S_L by $\bar{S}_L$, we denote the complementary pattern $\{(i,j) : (i,j) \notin S_L\}$. For a sparsity pattern S_L satisfying (2.11), we set

$$J_k = \{j \leqq k : (k,j) \notin S_L\}, \qquad k = 1, \ldots, n. \tag{3.1}$$

LEMMA 3.1. *Let S_L satisfy* (2.11),

$$J_k \times J_k \subseteq \bar{S}_L \cup \bar{S}_L^T, \tag{3.2}$$

and $S_U = S_L^T$. Then for any $n \times n$ matrix C and nonsingular matrices L and U of the sparsity patterns S_L and S_U, respectively, the conditions

$$\begin{aligned} (LC)_{ij} &= 0, \qquad (i,j) \in \bar{S}_L, \\ (CU)_{ij} &= 0, \qquad (i,j) \in \bar{S}_L^T, \end{aligned} \tag{3.3}$$

imply that

$$C_{ij} = 0 \quad \textit{if}\,(i,j) \in \bar{S}_L \cup \bar{S}_L^T.$$

Remark 3.1. Condition (3.2) can be equivalently rewritten in the form

(3.2′) $(k,i) \in \bar{S}_L^T$ and $(i,j) \in \bar{S}_L$ imply that $(k,j) \in \bar{S}_L \cup \bar{S}_L^T$ for any i, j, k,

implying that the matrix $P(\bar{S}_L + \bar{S}_L^T)P^T$, where P is the $n \times n$ matrix of the maximum length permutation $\pi : \pi(i) = n - i + 1$, is a perfect elimination matrix (as defined in [14]). In particular, condition (3.2) (as well as the perfect elimination condition) is satisfied if S_L is the sparsity pattern of a staircase or a banded lower triangular matrix.

Proof. We prove the lemma by induction on i. If $i = 1$ then $J_i = \{1\}$, and by (3.3), $L_{11}C_{11} = 0$ implying (since L is nonsingular) that $C_{11} = 0$. Assuming that $C_{rs} = 0$ if r, $s \leqq i - 1$ and $(r,s) \in \bar{S}_L \cup \bar{S}_L^T$, we now show that $C_{rs} = 0$ if $(r,s) \in \bar{S}_L \cup \bar{S}_L^T$ for r, $s \leqq i$. Indeed, by (3.3) for $j \in J_i \setminus \{i\}$

$$0 = (LC)_{ij} = L_{ii}C_{ij} + \sum_{k \in J_i \setminus \{i\}} L_{ik}C_{kj}, \tag{3.4}$$

and since the conditions $j \in J_i$ and $k \in J_i$ imply, in view of (3.2), that $(k,j) \in \bar{S}_L \cup \bar{S}_L^T$, by induction the sum in (3.4) is zero. Therefore, it follows from (3.4) that $C_{ij} = 0$ for $j \in J_i \setminus \{i\}$. The relations $C_{ji} = 0$, $j \in J_i \setminus \{i\}$, are derived in a similar way. Finally, the last relations and (3.3) imply that

$$0 = (LC)_{ii} = L_{ii}C_{ii} + \sum_{k \in J_i \setminus \{i\}} L_{ik}C_{ki} = L_{ii}C_{ii},$$

which shows that $C_{ii} = 0$. This completes the proof of Lemma 3.1. □

THEOREM 3.1. *Let A be a positive definite matrix and a sparsity pattern S_L satisfy* (2.11) *and* (3.2). *Then for G_L, G_U, and D defined by* (2.12) *and* (2.14) *with $S_U = S_L^T$, the following equalities hold*:

$$\begin{aligned} (G_L^{-1}DG_U^{-1})_{ij} &= A_{ij}, \qquad (i,j) \in \bar{S}_L \cup \bar{S}_L^T, \\ (G_U D^{-1} G_L)_{ij} &= 0, \qquad (i,j) \notin \bar{S}_L \cup \bar{S}_L^T. \end{aligned}$$

Proof. We note first that under the hypotheses of the theorem, the matrices G_L and G_U are nonsingular by Lemma 2.1, and

$$(DG_U^{-1})_{ii} = (G_L^{-1}D)_{ii} = 1, \qquad i = 1, \ldots, n, \tag{3.5}$$

by Corollary 2.1. Set $C := G_L^{-1} D G_U^{-1} - A$. Then by (2.12) and (3.5),

$$(G_L C)_{ij} = (D G_U^{-1} - G_L A)_{ij} = 0 \quad \text{if}\,(i,j) \in \bar{S}_L$$

and similarly,

$$(C G_U)_{ij} = (G_L^{-1} D - A G_U)_{ij} = 0 \quad \text{if}\,(i,j) \in \bar{S}_L^T.$$

Thus by Lemma 3.1, $C_{ij} = (G_L^{-1} D G_U^{-1})_{ij} - A_{ij} = 0$ if $(i,j) \in \bar{S}_L \cup \bar{S}_L^T$. Let us now prove that $(G_U D^{-1} G_L)_{ij} = 0$ for $(i,j) \notin \bar{S}_L \cup \bar{S}_L^T$. Since

$$(G_U D^{-1} G_L)_{ij} = \sum_{(k:\, i \in J_k) \,\&\, (j \in J_k)} (G_u)_{ik} (D^{-1})_{kk} (G_L)_{kj},$$

and since by (3.2) the simultaneous conditions $i \in J_k$ and $j \in J_k$ imply that $(i,j) \in \bar{S}_L \cup \bar{S}_L^T$, we conclude that $(G_U D^{-1} G_L)$ can be nonzero only if $(i,j) \in \bar{S}_L \cup \bar{S}_L^T$ and Theorem 3.1 is proved. □

Theorem 3.1 shows that for a positive definite matrix A there exists a structurally symmetric sparse factorized approximate inverse $G_U D^{-1} G_L$ with the sparsity pattern of $G_L + G_U$, whose inverse coincides with A up to positions at which the triangular factors G_L and G_U are allowed to have nonzero entries. Note that in the case when A is symmetric positive definite and $\bar{S}_L$ is banded, the statement of Theorem 3.1 was proved in [3] using a recursive definition of G_L and G_U instead of (2.12).

The last result of this section is not related to the preconditioning problem but is important by itself.

THEOREM 3.2. *Let A be an $n \times n$ positive definite matrix and*

$$(A^{-1})_{ij} = 0 \quad \textit{if}\,(i,j) \notin \bar{S},$$

where the index set $\bar{S}$ is symmetric,

$$\{(i,i) : 1 \leqq i \leqq n\} \subseteq \bar{S} \subseteq \{(i,j) : 1 \leqq i,j \leqq n\},$$

and satisfies the condition

$$\textit{if}\,(k,i) \in \bar{S},\, k \leqq i \quad \textit{and} \quad (i,j) \in \bar{S},\, i \geqq j, \quad \textit{then}\,(k,j) \in \bar{S}.$$

Then the triangular factorization $A^{-1} = G_U D^{-1} G_L$ of the inverse to A can be computed in parallel using the entries A_{ij} at positions $(i,j) \in \bar{S}$ via relations (2.12) *and* (2.14) *with*

$$S_L = \{(i,j) : \textit{either } i < j \textit{ or } i > j \textit{ and } (i,j) \notin \bar{S}\}$$

and $S_U = S_L^T$.

Proof. In view of the stated connection of (3.2′) with the perfect elimination condition, the matrix A^{-1} has the factorization $A^{-1} = UL$ with the triangular factors U and L of the sparsity patterns S_L^T and S_L, respectively. Since $LA = U^{-1}$ and $AU = L^{-1}$ we have, in particular,

$$(LA)_{ii} \neq 0, \qquad (AU)_{ii} \neq 0,$$

and

$$(LA)_{ij} = 0,\, i \neq j,\, (i,j) \notin S_L, \qquad (AU)_{ij} = 0,\, i \neq j,\, (i,j) \notin S_L^T.$$

Therefore, for some nonsingular diagonal matrices D_1 and D_2, we have

$$L = D_1 G_L \quad \text{and} \quad U = G_U D_2,$$

and it remains to show that $D^{-1} = D_2 D_1$. We have

$$G_L A G_U D^{-1} = D_1^{-1} L A U D_2^{-1} D^{-1} = D_1^{-1} D_2^{-1} D^{-1}. \tag{3.6}$$

On the other hand,

$$\begin{aligned}(G_L A G_U D^{-1})_{ii} &= \sum_{k:(i,k)\in S_L \,\&\, (k,i)\notin S_L^T} (G_L A)_{ik}(G_U D^{-1})_{ki} + (G_L A)_{ii}(G_U D^{-1})_{ii} \\ &= (G_L A)_{ii}(G_U D^{-1})_{ii} = (G_U D^{-1})_{ii},\end{aligned}$$

and thus, by Corollary 2.1, $(G_L A G_U D^{-1})_{ii} = 1$. In view of (3.6), this means that $D^{-1} = D_2 D_1$ and the proof is completed. □

4. Convergence of factorized sparse approximate inverse preconditionings. In this section we prove the convergence results for preconditionings of type (2.13) for M-, H-, and block H-matrices. Consider first the case of M-matrices.

LEMMA 4.1 [6], [7]. *Let A be an $n \times n$ M-matrix and $S \subseteq \{(i,j) : 1 \leqq i \neq j \leqq n\}$ be a sparsity pattern. Then for G determined by* (1.6), *the matrix GA is an M-matrix, $G \geqq 0$ is nonsingular, and the splitting*

$$A = G^{-1} - R \tag{4.1}$$

is convergent, i.e., $\rho(I - GA) < 1$.

Proof. First let us show that G is nonnegative and has no zero rows. Indeed, as mentioned in § 2, the ith row of $G = (g_{ij})$ coincides with the ith row of the inverse to $A^{(i)}(S)$ (see (2.8)), and since $A^{(i)}(S)$ is an M-matrix, its inverse is nonnegative and has no zero rows.

We show next that the off-diagonal entries of GA are nonpositive. For $(i,j) \notin S$, $i \neq j$, it is true by construction. For $(i,j) \in S$ we have

$$(GA)_{ij} = \sum_{k:(i,k)\notin S} g_{ik} a_{kj} \leqq 0$$

since $g_{ik} \geqq 0$, and the conditions $(i,j) \in S$ and $(i,k) \notin S$ imply that $k \neq j$ and therefore $a_{kj} \leqq 0$.

To prove that GA is an M-matrix, it is now sufficient [1] to provide a positive vector v such that GAv is also positive. But A being an M-matrix there exists a vector $v > 0$ such that $Av > 0$ [1] and hence $GAv > 0$, since all rows of G are nonnegative and nonzero. Thus GA is an M-matrix implying, in particular, that G is nonsingular.

Finally, the matrices G and $GR = I - GA$ being nonnegative, the splitting (4.1) is weak-regular and thus convergent [1], which completes the proof of the lemma. □

LEMMA 4.2. *Let L and U be nonnegative, nonsingular, respectively, lower and upper triangular matrices such that LA and AU are M-matrices. Then LAU is also an M-matrix.*

Proof. We show first that off-diagonal entries of LAU are nonpositive. Indeed, let $i < j$. Then

$$(LAU)_{ij} = \sum_{k\leqq i} L_{ik}(AU)_{kj} \leqq 0,$$

since the conditions $k \leqq i$ and $i < j$ imply that $k \neq j$ and therefore, $(AU)_{kj} \leqq 0$ as an off-diagonal entry of an M-matrix AU. Similarly, for $j < i$, we see that

$$(LAU)_{ij} = \sum_{k\leqq j} (LA)_{ik} U_{kj} \leqq 0.$$

To show that LAU is an M-matrix, it is now sufficient [1] to provide a positive vector v such that $LAUv > 0$. Since, by assumption, AU is an M-matrix, there exists [1] a vector $v > 0$ such that $AUv > 0$ and therefore, L being nonnegative and nonsingular, $LAUv$ is also positive. The lemma is thus proved. □

THEOREM 4.1. *Let A be an M-matrix, the sparsity patterns S_L and S_U satisfy* (2.11), *and G_L and G_U be defined by* (2.12). *Then the matrix G_LAG_U is also an M-matrix and the splitting*

$$G_LAG_U = D - R, \qquad D = \operatorname{diag}(G_LAG_U), \tag{4.2}$$

is convergent.

Proof. Applying Lemma 4.1 to A and G_L and to A^T and G_U^T, we obtain that G_LA and AG_U are M-matrices, while G_L and G_U are nonsingular and nonnegative. Thus, by Lemma 4.2, G_LAG_U is an M-matrix and splitting (4.2) is convergent as a regular splitting of an M-matrix [1]. □

Next we prove convergence of splitting (4.2) for an H-matrix A. To this end we set $\tilde{A} := \mathcal{M}(A)$, where the comparison matrix $\mathcal{M}(A)$ (which is an M-matrix by the definition of an H-matrix) is defined by the relations

$$[\mathcal{M}(A)]_{ij} = \begin{cases} |a_{ii}|, & i = j, \\ -|a_{ij}|, & i \neq j. \end{cases}$$

For chosen sparsity patterns S_L and S_U in parallel with the preconditioning $A \to G_LAG_UD^{-1}$, we consider the similar preconditioning $\tilde{A} \to \tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}^{-1}$, where both preconditionings are constructed using (2.12) and (2.14).

LEMMA 4.3 [7]. *Let A be an H-matrix and $\tilde{A} = \mathcal{M}(A)$. Then for any sparsity pattern $S \subseteq \{(i,j) : 1 \leqq i \neq j \leqq n\}$, the matrices $G = (g_{ij})$ and $\tilde{G} = (\tilde{g}_{ij})$ such that*

$$\begin{aligned} g_{ij} = \tilde{g}_{ij} &= 0 && \text{if } (i,j) \in S, \\ (GA)_{ij} = (\tilde{G}\tilde{A})_{ij} &= \delta_{ij} && \text{if } (i,j) \notin S \end{aligned} \tag{4.3}$$

satisfy the inequality

$$|G| \leqq \tilde{G}.$$

Proof. Since the ith rows of G and $\tilde{G}$ coincide, respectively, with the ith rows of $(A^{(i)}(S))^{-1}$ and $(\tilde{A}^{(i)}(S))^{-1}$ defined according to (2.8), and since $\tilde{A}^{(i)}(S) = \mathcal{M}(A^{(i)}(S))$, the required inequality $|G| \leqq \tilde{G}$ follows from the Ostrowski theorem [12]. The lemma is thus proved. □

LEMMA 4.4 [7]. *Under the hypotheses of Lemma* 4.3, $|GA| \leqq |\tilde{G}\tilde{A}|$, *and GA is an H-matrix.*

Proof. Note that by (4.3) for $(i,j) \notin S$, we have $(GA)_{ij} = (\tilde{G}\tilde{A})_{ij}$. So let $(i,j) \in S$. Then

$$|GA|_{ij} = \left| \sum_{k:(i,k)\notin S} g_{ik}a_{kj} \right| \leqq \sum_{k:(i,k)\notin S} |g_{ik}a_{kj}|. \tag{4.4}$$

Since conditions $(i,j) \in S$ and $(i,k) \notin S$ imply that $k \neq j$, the right-hand side of (4.4) does not involve diagonal entries of A. Therefore, taking into account that $\tilde{A}$ is an M-matrix, it follows from (4.4) and Lemma 4.3 that

$$|GA|_{ij} \leqq \sum_{k:(i,k)\notin S} \tilde{g}_{ik}|\tilde{a}_{kj}| = |\tilde{G}\tilde{A}|_{ij},$$

implying that $\mathcal{M}(GA) \geqq \tilde{G}\tilde{A}$. Since $\tilde{G}\tilde{A}$ is an M-matrix by Lemma 4.1, GA is an H-matrix, which completes the proof. □

Note that when $S = \{(i,j) : i < j\}$ the matrices G and $\tilde{G}$ defined in Lemma 4.3 are lower triangular, while the matrices $U := GA$ and $\tilde{U} := \tilde{G}\tilde{A}$ are upper unitriangular. By Lemma 4.4 the latter matrices satisfy the inequality $|U| \leqq |\tilde{U}|$. Taking into account that $\tilde{U}$ is an M-matrix by Lemma 4.1, we rewrite the last inequality in the form $\mathcal{M}(U) \geqq \tilde{U}$, which implies, by the Ostrowski theorem [12], that

$$|U^{-1}| \leqq \mathcal{M}(U^{-1}) \leqq \tilde{U}^{-1}.$$

We have thus proved the following lemma.

LEMMA 4.5. *Let A be an H-matrix and $\tilde{A} = \mathcal{M}(A)$. If $A = LU$ and $\tilde{A} = \tilde{L}\tilde{U}$ are the triangular decompositions of A and $\tilde{A}$ such that U and $\tilde{U}$ are upper unitriangular, then*

$$|U^{-1}| \leqq \mathcal{M}(U^{-1}) \leqq \tilde{U}^{-1}.$$

THEOREM 4.2. *Let A be an H-matrix and G_L and G_U be defined by* (2.12) *for some sparsity patterns S_L and S_U satisfying* (2.11). *Then G_LAG_U is also an H-matrix and the splitting*

$$G_LAG_U = D - R, \qquad D = \operatorname{diag}(G_LAG_U),$$

is convergent, i.e., $\rho(I - G_LAG_UD^{-1}) < 1$.

Proof. Note first that by construction, the subvector of the ith column of G_U made up of the entries at the positions $(k, i) \notin S_U$ coincides with the last column of the inverse to the principal submatrix $A[I_i, I_i]$ of A, where $I_i = \{k \leqq i : (k, i) \notin S_U\}$. This implies that the corresponding subvector of $G_UD_U^{-1}$, where $D_U = \operatorname{diag}(G_U)$, coincides with the last column of the inverse to the upper unitriangular factor of the LU decomposition of $A[I_i, I_i]$. For $\tilde{A} = \mathcal{M}(A)$, define $\tilde{G}_U$ by the relations

$$(\tilde{G}_U)_{ij} = 0, \qquad (i,j) \in S_U,$$

$$(\tilde{G}_U\tilde{A})_{ij} = \delta_{ij}, \qquad (i,j) \notin S_U.$$

Then applying Lemma 4.5 we obtain

$$|G_UD_U^{-1}| \leqq \tilde{G}_U\tilde{D}_U^{-1}, \tag{4.5}$$

where $\tilde{D}_U = \operatorname{diag}(\tilde{G}_U)$.

We show next that D is nonsingular and

$$|D_UD^{-1}| \leqq \tilde{D}_U\tilde{D}^{-1}, \tag{4.6}$$

where $\tilde{D} = \operatorname{diag}(\tilde{G}_L\tilde{A}\tilde{G}_U)$ and $\tilde{G}_L$ is the corresponding left preconditioner for $\tilde{A}$. Indeed, since $(G_LA)_{ii} = 1$, we have

$$(DD_U^{-1})_{ii} = 1 + \sum_{k<i} (G_LA)_{ik}(G_UD_U^{-1})_{ki},$$

implying, in view of Lemma 4.4, inequality (4.5), and Lemma 4.1, that

$$|DD_U^{-1}|_{ii} \geqq 1 - \sum_{k<i} |G_LA|_{ik}|G_UD_U^{-1}|_{ki} \geqq 1 - \sum_{k<i} |\tilde{G}_L\tilde{A}|_{ik}|\tilde{G}_U\tilde{D}_U^{-1}|_{ki} = (\tilde{D}\tilde{D}_U^{-1})_{ii} > 0,$$

and (4.6) follows. Let us now prove the inequality

$$|G_LAG_UD_U^{-1}|_{ij} \leqq |\tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}_U^{-1}|_{ij}, \qquad i \neq j. \tag{4.7}$$

First let $i > j$. Then by Lemma 4.4 and (4.5),

$$\begin{aligned} |G_LAG_UD_U^{-1}|_{ij} &= \left|\sum_{k\leqq j}(G_LA)_{ik}(G_UD_U^{-1})_{kj}\right| \\ &\leqq \sum_{k\leqq j}|\tilde{G}_L\tilde{A}|_{ik}|\tilde{G}_U\tilde{D}_U^{-1}|_{kj} \underset{(k\neq i)}{=} |\tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}_U^{-1}|_{ij}. \end{aligned}$$

Consider now the case $i < j$. To this end we show that

$$|AG_UD_U^{-1}|_{kj} \leqq |\tilde{A}\tilde{G}_U\tilde{D}_U^{-1}|_{kj}, \qquad k \neq j. \tag{4.8}$$

Indeed, if $(k, j) \notin S_U$, $k \neq j$, then both parts of (4.8) are zero, while if $(k, j) \in S_U$, then using (4.5) again, we see that

$$|AG_UD_U^{-1}|_{kj} = \left|\sum_{(r,j)\notin S_U} A_{kr}(G_UD_U^{-1})_{rj}\right| \underset{(k\neq r)}{\leqq} |\tilde{A}\tilde{G}_U\tilde{D}_U^{-1}|_{kj},$$

and (4.8) is thus established. Now for $i < j$, by Lemma 4.3 and inequality (4.8),

$$\begin{aligned} |G_LAG_UD_U^{-1}|_{ij} &= \left|\sum_{k\leqq i}(G_L)_{ik}(AG_UD_U^{-1})_{kj}\right| \\ &\overset{(k\neq j)}{\leqq} \sum_{k\leqq i}(\tilde{G}_L)_{ik}|\tilde{A}\tilde{G}_U\tilde{D}_U^{-1}|_{kj} = |\tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}_U^{-1}|_{ij}, \end{aligned}$$

which completes the proof of (4.7).

We are now able to show that

$$|I - G_LAG_UD^{-1}| \leqq I - \tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}^{-1}. \tag{4.9}$$

Indeed, for $i = j$, (4.9) is trivially fulfilled, while for $i \neq j$, by (4.7) and (4.6), we have

$$\begin{aligned} |I - G_LAG_UD^{-1}|_{ij} &= |G_LAG_UD^{-1}|_{ij} = |G_LAG_UD_U^{-1}|_{ij}|D_UD^{-1}|_{jj} \\ &\leqq |\tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}_U^{-1}|_{ij}(\tilde{D}_U\tilde{D}^{-1})_{jj} = |\tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}^{-1}|_{ij} \\ &= (I - \tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}^{-1})_{ij}. \end{aligned}$$

Finally, (4.9) implies that

$$\rho(I - G_LAG_UD^{-1}) \leqq \rho(|I - G_LAG_UD^{-1}|) \leqq \rho(I - \tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}^{-1}),$$

and since by Theorem 4.1, $\rho(I - \tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}^{-1}) < 1$, the required convergence result follows. Furthermore, since

$$\mathcal{M}(G_LAG_UD^{-1}) \geqq \tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}^{-1}$$

by (4.9), and $\tilde{G}_L\tilde{A}\tilde{G}_U\tilde{D}^{-1}$ is an M-matrix by Theorem 4.1, both $G_LAG_UD^{-1}$ and G_LAG_U are H-matrices. Theorem 4.2 is thus completely proved. □

We conclude this section by considering the case of block H-matrices. Let a matrix A be partitioned into blocks $A = (A_{ij})$, $1 \leqq i, j \leqq n$, in such a way that the diagonal blocks A_{ii} are all square and nonsingular. Following [13], A is said to be a block H-matrix (with respect to the chosen block partitioning) if its block comparison $n \times n$ matrix $\mathcal{M}_b(A)$ with the entries

$$(\mathcal{M}_b(A))_{ij} = \begin{cases} 1, & i = j \\ -\|A_{ii}^{-1}A_{ij}\|, & i \neq j \end{cases} \qquad 1 \leqq i, j \leqq n, \tag{4.10}$$

is an M-matrix. In (4.10), we use the notation

$$\|A_{ij}\| = \|A_{ij}\|_{ij} = \sup_{x\in\Omega_j, x\neq 0} \frac{\|A_{ij}x\|_{\Omega_i}}{\|x\|_{\Omega_j}}, \qquad 1 \leqq i, j \leqq n,$$

where Ω_i is the vector space on which the linear operator acts with the matrix A_{ii}, and $\|\cdot\|_{\Omega_i}$ is a vector norm on Ω_i.

We need the following well-known block counterpart of the Ostrowski theorem.

THEOREM 4.3 [13]. *A block H-matrix $A = (A_{ij})$, $1 \leqq i, j \leqq n$, is nonsingular and*

$$\|(A^{-1})_{ij}\| \leqq (\mathcal{M}_b^{-1}(A))_{ij} \|A_{jj}^{-1}\|, \qquad 1 \leqq i, j \leqq n.$$

For a block $n \times n$ matrix A and a sparsity pattern $S \subseteq \{(i,j) : 1 \leqq i \neq j \leqq n\}$, define the block $n \times n$ matrix $G = (G_{ij})$ of block sparsity pattern S by the following blockwise relations:

$$\begin{aligned} G_{ij} &= 0, && (i,j) \in S, \\ (GA)_{ij} &= I_i, && i = j, \\ (GA)_{ij} &= 0, && (i,j) \notin S, \quad i \neq j. \end{aligned} \tag{4.11}$$

Set $\tilde{A} = \mathcal{M}_b(A)$ and define, in parallel with G, the $n \times n$ matrix $\tilde{G} = (\tilde{g}_{ij})$ by the pointwise relations

$$\begin{aligned} \tilde{g}_{ij} &= 0, && (i,j) \in S, \\ (\tilde{G}\tilde{A})_{ij} &= \delta_{ij}, && (i,j) \notin S. \end{aligned} \tag{4.12}$$

Define also the block matrix $\hat{A} = (\hat{A}_{ij})$ by the relations

$$\hat{A}_{ij} = A_{ii}^{-1} A_{ij}, \qquad 1 \leqq i, j \leqq n,$$

and the block matrix $\hat{G} = (\hat{G}_{ij})$ by the relations

$$\begin{aligned} \hat{G}_{ij} &= 0, && (i,j) \in S, \\ (\hat{G}\hat{A})_{ij} &= I_i, && i = j, \\ (\hat{G}\hat{A})_{ij} &= 0, && (i,j) \notin S, \quad i \neq j. \end{aligned} \tag{4.13}$$

Obviously, when A is a block H-matrix, then $\hat{A}$ is well defined and $\mathcal{M}_b(\hat{A}) = \mathcal{M}_b(A)$, implying that $\hat{A}$ is also a block H-matrix.

The following three lemmata are the block counterparts of Lemmata 4.3–4.5.

LEMMA 4.6. *Let $A = (A_{ij})$, $1 \leqq i, j \leqq n$, be a block H-matrix and let $\tilde{A} = \mathcal{M}_b(A)$. Then for any sparsity pattern $S \subseteq \{(i,j) : 1 \leqq i \neq j \leqq n\}$, the matrices $\hat{G} = (\hat{G}_{ij})$ and $\tilde{G} = (\tilde{G}_{ij})$, defined according to* (4.13) *and* (4.12), *satisfy the relations*

$$\|\hat{G}_{ij}\| \leqq \tilde{g}_{ij}, \qquad 1 \leqq i, j \leqq n.$$

Proof. Apply Theorem 4.3 for the proof. □

LEMMA 4.7. *Under the hypotheses of Lemma* 4.6 *it holds that*

$$\|(GA)_{ij}\| \leqq |\tilde{G}\tilde{A}|_{ij}, \qquad 1 \leqq i, j \leqq n,$$

implying that GA is again a block H-matrix.

Proof. Note that $\hat{G} = G \cdot$ block diag (A), and thus

$$GA = \hat{G}\hat{A}. \tag{4.14}$$

For $(i, j) \in S$ we have, by Lemma 4.6 and equation (4.10),

$$\|(\hat{G}\hat{A})_{ij}\| \leqq \sum_{k:(i,k)\notin S} \|\hat{G}_{ik}\| \, \|\hat{A}_{kj}\| \leqq \sum_{k:(i,k)\notin S} \tilde{g}_{ik}|\tilde{a}_{kj}| \overset{(k\neq j)}{=} |\tilde{G}\tilde{A}|_{ij},$$

implying, in view of (4.14), that

$$\|(GA)_{ij}\| \leqq |\tilde{G}\tilde{A}|_{ij}.$$

The last inequality being satisfied for $(i, j) \notin S$ by construction and $\tilde{G}\tilde{A}$ being an M-matrix by Lemma 4.1, this completes the proof of Lemma 4.7. □

The proof of the following lemma is analogous to that of Lemma 4.5 and is omitted.

LEMMA 4.8. *Let $A = (A_{ij})$, $1 \leqq i, j \leqq n$, be a block H-matrix and let $A = LU$, where L is a lower block triangular matrix, while U is an upper block triangular matrix with unit diagonal blocks $U_{ii} = I_i$, $1 \leqq i \leqq n$. Similarly, for $\tilde{A} = \mathcal{M}_b(A)$, let $\tilde{A} = \tilde{L}\tilde{U}$, where $\tilde{L}$ is lower triangular and $\tilde{U}$ is upper unitriangular. Then*

$$\|(U^{-1})_{ij}\| \leqq (\mathcal{M}_b^{-1}(U))_{ij} \leqq (\tilde{U}^{-1})_{ij}, \qquad 1 \leqq i, j \leqq n.$$

Convergence of sparse factorized approximate inverse preconditionings for block H-matrices is established in the following.

THEOREM 4.4. *Let $A = (A_{ij})$, $1 \leqq i, j \leqq n$, be a block H-matrix and let, for some sparsity patterns S_L and S_U satisfying* (2.11), *the block triangular matrices G_L and G_U be defined by the following blockwise relations*:

$$\begin{array}{lll lll}
(G_L)_{ij} = 0, & (i,j) \in S_L; & & (G_U)_{ij} = 0, & (i,j) \in S_U; & \\
(G_LA)_{ij} = I_i, & i = j; & & (AG_U)_{ij} = I_i, & i = j; & \\
(G_LA)_{ij} = 0, & (i,j) \notin S_L, & i \neq j; & (AG_U)_{ij} = 0, & (i,j) \notin S_U, & i \neq j.
\end{array}$$

Then G_LAG_U is a block H-matrix and the splitting

$$G_LAG_U = D - R, \qquad D = \text{block diag}\,(G_LAG_U),$$

is convergent.

The proof of Theorem 4.4 follows that of Theorem 4.2 with $\tilde{A} = \mathcal{M}_b(A)$, and so we omit it except for the counterpart of (4.6):

$$\|(D_UD^{-1})_{ii}\| \leqq (\tilde{D}_U\tilde{D}^{-1})_{ii}, \qquad 1 \leqq i \leqq n, \tag{4.15}$$

which is proved as follows. Recall first that for a nonsingular matrix $I - P$ with $\|P\| < 1$, we have

$$\|(I - P)^{-1}\| \leqq \frac{1}{1 - \|P\|}. \tag{4.16}$$

Fix an i, $1 \leqq i \leqq n$, and set

$$P := - \sum_{k:k<i} (G_LA)_{ik}(G_UD_U^{-1})_{ki}.$$

Then the required inequality (4.15) follows from (4.16) since

$$(DD_U^{-1})_{ii} = I - P$$

and (by Lemmata 4.7 and 4.8)

$$\begin{aligned}\|P\| &\leqq \sum_{k:k<i} \|(G_L A)_{ik}\| \, \|(G_U D_U^{-1})_{ki}\| \\ &\leqq \sum_{k:k<i} |\tilde{G}_L \tilde{A}|_{ik} |\tilde{G}_U \tilde{D}_U^{-1}|_{ki} = 1 - (\tilde{D}\tilde{D}_U^{-1})_{ii} < 1. \qquad \square\end{aligned}$$

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences,* Academic Press, New York, 1979.

[2] M. W. Benson and P. O. Frederickson, *Iterative solution of large sparse linear systems arising in certain multidimensional approximation problems*, Utilitas Math., 22 (1982), pp. 127–140.

[3] P. Dewilde and E. F. Deprettere, *Approximate inversion of positive matrices with applications to modelling*, in Modelling, Robustness, and Sensitivity Reduction in Control Systems, R. F. Curtain, ed., NATO ASI Series, F34 (1987), pp. 211–238.

[4] G. E. Forsythe and E. G. Strauss, *On the best conditioned matrices*, Proc. Amer. Math. Soc., 6 (1955), pp. 340–345.

[5] I. E. Kaporin, *An alternative approach to estimating the convergence rate of the* CG *method*, in Numerical Methods and Software, Yu. A. Kuznetsov, ed., Dept. of Numerical Mathematics, USSR Academy of Sciences, Moscow, 1990, pp. 55–72. (In Russian.)

[6] L. Yu. Kolotilina, *A family of explicit preconditionings for sparse linear algebraic systems*, Preprint of LOMI P-8-86, V. A. Steklov Mathematical Inst., USSR Academy of Sciences, Leningrad, 1986. (In Russian.)

[7] ———, *Explicit preconditioning of H-matrices*, in Numerical Analysis and Mathematical Modelling, Yu. A. Kuznetsov, ed., Dept. of Numerical Mathematics, USSR Academy of Sciences, Moscow, 1989, pp. 97–108. (In Russian.)

[8] L. Yu. Kolotilina and A. Yu. Yeremin, *On a family of two-level preconditionings of the incomplete block factorization type*, Soviet J. Numer. Anal. Math. Model., 1 (1986), pp. 293–320.

[9] ———, *Incomplete block factorization methods for complex structure matrices*, Zapiski Nauchn. Semin. LOMI, 159 (1987), pp. 5–22. (In Russian.) J. Soviet Math., 47 (1989), pp. 2821–2834.

[10] ———, *Factorized sparse approximate inverse preconditionings* II: *Solution of* 3D FE *systems on massively parallel computers*, manuscript.

[11] L. Yu. Kolotilina, A. A. Nikishin, and A. Yu. Yeremin, *Factorized sparse approximate inverse preconditionings* III: *Iterative construction of preconditioners for* 3D FE *problems on massively parallel computers*, manuscript.

[12] A. M. Ostrowski, *Über die Determinanten mit überwiegender Hauptdiagonale*, Comment. Math. Helv., 10 (1937), pp. 69–96.

[13] F. Robert, *Blocs-H-matrices et convergence des méthodes itératives classiques par blocs*, Linear Algebra Appl., 2 (1969), pp. 223–265.

[14] D. J. Rose, *A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations*, in Graph Theory and Computing, Academic Press, New York, 1972, pp. 183–217.

[15] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 59–73, January 1993

005

ESTIMATION OF THE OPTIMUM RELAXATION FACTORS IN PARTIAL FACTORIZATION ITERATIVE METHODS*

Z. I. WOŹNICKI†

Abstract. A new approach is proposed for estimating the optimum relaxation factors in both single and double successive overrelaxation (SOR) processes used for the acceleration of convergence in the partial factorization methods called *two-sweep iterative methods*, and introduced for the solution of large, sparse, linear systems appearing in the difference approximations to the elliptic partial differential equations.

Key words. linear equation systems, sparse matrices, partial factorizations, (single and double splitting) iterative methods, eigenvalues, optimum relaxation factor estimation

AMS(MOS) subject classifications. 65F10, 65F15, 65F50, 15A23

1. Introduction. The concept of the estimation of optimum relaxation factors in partial factorization iterative methods is derived from the analysis of the iteration matrix spectral radius against the relaxation factor for the model based on the Dirichlet problem.

All considerations are referred to the solution of linear equation system

$$A\phi = c, \tag{1}$$

in which the $n \times n$ nonsingular matrix A is defined by the following decomposition:

$$A = K - L - U, \tag{2}$$

where K, L, and U are diagonal, strictly lower triangular, and strictly upper triangular matrices, respectively. It is assumed that K is a nonsingular matrix.

2. Single splitting iterative schemes. The iterative solution of (1) can be expressed in the form

$$M\phi^{(t+1)} = N\phi^{(t)} + c, \qquad t \geqq 0, \tag{3}$$

where $\phi^{(t)}$ denotes the successive iterates and

$$A = M - N \tag{4}$$

represents the *single splitting* of A as the classical splitting of A in the theory of iterative methods. The above iterative scheme is convergent to the unique solution

$$\phi = A^{-1}c \tag{5}$$

for each $\phi^{(0)}$ if and only if M is a nonsingular matrix and the corresponding *iteration matrix*

$$\mathcal{G} = M^{-1}N \tag{6}$$

has the *spectral radius*

$$\rho(\mathcal{G}) = \max |\lambda_i| < 1, \qquad 1 \leqq i \leqq n, \tag{7}$$

where λ_i's are eigenvalues of $\mathcal{G}$. Equation (3) can be written in the equivalent form

$$\phi^{(t+1)} = \mathcal{G}\phi^{(t)} + M^{-1}c, \qquad t \geqq 0. \tag{8}$$

* Received by the editors July 20, 1990; accepted for publication (in revised form) July 30, 1991. This paper previously appeared in the Proceedings of the International Conference on the Physics of Reactors: Operation, Design, and Computation, PHYSOR 90, Marseille, France, April 23–27, 1990, pp. 173–176.

† Institute of Atomic Energy, 05-400 Otwock-Swierk, Poland (r05zw@cx1.cyf.gov.pl).

The spectral radius of the iteration matrix plays a fundamental role in the convergence analysis of iterative methods. The quantity

$$R_\infty(\mathscr{G}) = -\ln \rho(\mathscr{G}), \tag{9}$$

called the *asymptotic rate of convergence*, increases as the value of $\rho(\mathscr{G})$ is decreasing. Thus, in order to maximize the rate of convergence, the matrices M and N should be chosen in such a way that the spectral radius $\rho(\mathscr{G})$ is as small as possible. In practice, we might expect that once M is closer to A the method will converge faster.

The idea of factorization methods consists in expressing the matrix M as the product of nonsingular matrices chosen in such a way that they are easy to obtain and relatively easily invertible. Then $N = M - A$ can be considered as the *residual matrix* for the assumed factorization of M and, when N exists as a nonzero matrix, there is the *partial factorization* of A and the solution has the iterative nature. In the case when N becomes the null matrix there is the *strict factorization* of A and the solution of (1) is obtained by means of the direct method equivalent to Gaussian elimination.

Introducing additional strictly lower triangular and strictly upper triangular matrices H and Q, respectively, then with the definition of $A = K - L - U$, we can assume the following factorization

$$M = [I - (L + H)D^{-1}]D[I - D^{-1}(U + Q)], \tag{10}$$

where D is assumed to be a nonsingular diagonal matrix defined by the following implicit relation

$$D = K - \text{diag}\,\{(L + H)D^{-1}(U + Q)\}, \tag{11}$$

and, as can be easily verified,

$$N = \text{offdiag}\{(L + H)D^{-1}(U + Q)\} - H - Q, \tag{12}$$

where the notation diag $\{B\}$ denotes the diagonal matrix with diagonal entries identical with those of B and offdiag $\{B\} = B -$ diag $\{B\}$.

The iterative method can be written as follows:

$$\phi^{(t+1)} = \mathscr{F}\phi^{(t)} + M^{-1}c, \qquad t \geqq 0 \tag{13}$$

and

$$\mathscr{F} = M^{-1}N = [I - D^{-1}(U + Q)]^{-1}D^{-1}[I - (L + H)D^{-1}]^{-1}N. \tag{14}$$

Since $I - (L + H)D^{-1}$ and $I - D^{-1}(U + Q)$ are nonsingular lower and upper triangular matrices, respectively, this method can be easily implemented by using the *two-sweep procedure*.

Let us multiply (13) on the left by $I - D^{-1}(U + Q)$ and shift the term $D^{-1}(U + Q)\phi^{(t+1)}$ to the right-hand side of the equation

$$\phi^{(t+1)} = D^{-1}\{(U + Q)\phi^{(t+1)} + [I - (L + H)D^{-1}]^{-1}(N\phi^{(t)} + c)\}.$$

Denoting by

$$\beta^{(t+1)} = [I - (L + H)D^{-1}]^{-1}(N\phi^{(t)} + c)$$

and again multiplying this expression on the left by $I - (L + H)D^{-1}$, we finally obtain

$$\begin{aligned} \beta^{(t+1)} &= (L + H)D^{-1}\beta^{(t+1)} + N\phi^{(t)} + c, \\ \phi^{(t+1)} &= D^{-1}[(U + Q)\phi^{(t+1)} + \beta^{(t+1)}], \qquad t \geqq 0. \end{aligned} \tag{15}$$

Since $(L + H)D^{-1}$ and $D^{-1}(U + Q)$ are strictly lower triangular and strictly upper triangular matrices, respectively, the components of $\beta^{(t+1)}$ can be calculated recursively for increasing indices in the *forward elimination sweep*, and components of $\phi^{(t+1)}$ can be calculated recursively for decreasing indices in the *backward substitution sweep*. Equations (15) represent the general form of a broad class of methods called the *two-sweep iterative methods* and each of them is uniquely defined by choice of the matrices H and Q. The matrix $\mathcal{F}$ is called the *two-sweep iteration matrix*. The classification of factorization methods from the viewpoint of the choice of H and Q matrices is given in [5] and [6].

Let us restrict our attention to the iterative schemes defined by the following choice of matrices H and Q, including such classical schemes as the Jacobi and Gauss–Seidel methods.

1. *The Jacobi method.*

$$\begin{aligned} &H = -L, \quad Q = -U, \quad D_J = K, \\ &M_J = K, \qquad N_J = L + U, \\ &\mathcal{B}_1 = M_J^{-1}N_J = K^{-1}(L + U). \end{aligned}$$

2. *The Gauss–Seidel method.*

$$\begin{aligned} &H = -L, \quad Q = 0, \quad D_G = K, \\ &M_G = K(I - K^{-1}U), \qquad N_G = L, \\ &\mathcal{L}_1 = M_G^{-1}N_G = (I - K^{-1}U)^{-1}K^{-1}L. \end{aligned}$$

3. *The EWA method.*

$$\begin{aligned} &H = Q = 0, \qquad D_E = K - \operatorname{diag}\{LD_E^{-1}U\}, \\ &M_E = (I - LD_E^{-1})D_E(I - D_{\mathrm{E}}^{-1}U), \qquad N_E = \operatorname{offdiag}\{LD_E^{-1}U\}, \\ &\mathcal{E}_1 = M_E^{-1}N_E = (I - D_E^{-1}U)^{-1}D_E^{-1}(I - LD_E^{-1})^{-1}N_E. \end{aligned}$$

4. *The AGA method.*

$$\begin{aligned} &H = H_{\mathrm{A}}, \quad Q = Q_A, \quad D_A = K - \operatorname{diag}\{(L + H_A)D_A^{-1}(U + Q_A)\}, \\ &M_A = [I - (L + H_A)D_A^{-1}]D_A[I - D_A^{-1}(U + Q_A)], \\ &N_A = \operatorname{offdiag}\{(L + H_A)D_A^{-1}(U + Q_A)\} - H_A - Q_A, \\ &\mathcal{A}_1 = M_A^{-1}N_A = [I - D_A^{-1}(U + Q_A)]^{-1}D_{\mathrm{A}}^{-1}[I - (L + H_A)D_A^{-1}]^{-1}N_A. \end{aligned}$$

In the literature [7], [11], the Gauss–Seidel method is usually described by the iteration matrix $\mathcal{L}_1 = (I - K^{-1}L)K^{-1}U$, which corresponds to the case when $H = 0$ and $Q = -U$. Since in the two-sweep approach we used, the solution vector $\phi^{(t+1)}$ is calculated in the backward sweep, then to keep the consistent definitions of particular methods, the Gauss–Seidel method considered here is defined in reverse order.

The algorithm of the EWA method was implemented in the EWA-II code [10].

The AGA method represents a broad class of algorithms and each is defined by the choice of the matrices H_A and Q_A such that $H_A + Q_A \neq 0$, and the nonzero entries of N_A do not coincide in location with the nonzero entries of $L + H_A + U + Q_A$. The assumption that H_A and Q_A are strictly lower triangular and strictly upper triangular matrices, respectively, allows us to explicitly determine the values of nonzero entries in H_A and Q_A

for an arbitrary pattern of the location of nonzero entries in these matrices, directly from the implicit form of $(L + H_A)D_A^{-1}(U + Q_A)$. In other words, all nonzero entries of H_A, Q_A (and consequently, D_A), and N_A can be computed successively in the simple algorithm by means of recursive formulae for each pair of increasing entry indices, $1 \leqq i, j \leqq n$, either row-by-row or column-by-column [1], [2].

When A is an irreducibly diagonally dominant matrix and $A^{-1} > 0$, the following inequality has been proven [1], [2].

$$0 < \rho(\mathscr{B}_1) < \rho(\mathscr{L}_1) < \rho(\mathscr{E}_1) < \rho(\mathscr{A}_1) < 1. \tag{16}$$

In order to accelerate the convergence in the two-sweep iterative methods, the following modification of the two-sweep equations (15) is used

$$\begin{aligned} \beta^{(t+1)} &= \omega_\beta[(L + H)D^{-1}\beta^{(t+1)} + N\phi^{(t)} + c] - (\omega_\beta - 1)\beta^{(t)}, \\ \phi^{(t+1)} &= D^{-1}[(U + Q)\phi^{(t+1)} + \beta^{(t+1)}], \qquad t \geqq 0. \end{aligned} \tag{17}$$

This defines the method called the *forward successive overrelaxation two-sweep iterative method* (*forward* SOR) and ω_β is the *relaxation factor*. By the elimination of the vector β, we obtain the following equation

$$M_{\omega_\beta}\phi^{(t+1)} = N_{\omega_\beta}\phi^{(t)} + c, \tag{18}$$

where

$$M_{\omega_\beta} = (1/\omega_\beta)[I - \omega_\beta(L + H)D^{-1}]D[I - D^{-1}(U + Q)], \tag{19}$$

$$N_{\omega_\beta} = (1/\omega_\beta)\{\omega_\beta N - (\omega_\beta - 1)D[I - D^{-1}(U + Q)]\}, \tag{20}$$

and

$$\begin{aligned} \mathscr{F}_{\omega_\beta} &= [I - D^{-1}(U + Q)]^{-1}D^{-1}[I - \omega_\beta(L + H)D^{-1}]^{-1} \\ &\quad \times \{\omega_\beta N - (\omega_\beta - 1)D[-D^{-1}(U + Q)]\}. \end{aligned} \tag{21}$$

Relating this acceleration procedure to the iteration methods defined by the particular choices of the matrices H and Q, we have the following methods.

5. *The Jacobi forward* SOR.

$$M_{J\omega_\beta} = (1/\omega_\beta)K,$$

$$N_{J\omega_\beta} = (1/\omega_\beta)[\omega_\beta(L + U) - (\omega_\beta - 1)K],$$

$$\mathscr{B}_{\omega_\beta} = M_{J\omega_\beta}^{-1}N_{J\omega_\beta} = \omega_\beta\mathscr{B}_1 - (\omega_\beta - 1)I.$$

6. *The Gauss-Seidel forward* SOR.

$$M_{G\omega_\beta} = (1/\omega_\beta)K(I - K^{-1}U),$$

$$N_{G\omega_\beta} = (1/\omega_\beta)[\omega_\beta L - (\omega_\beta - 1)K(I - K^{-1}U)],$$

$$\mathscr{L}_{\omega_\beta} = M_{G\omega_\beta}^{-1}N_{G\omega_\beta} = \omega_\beta\mathscr{L}_1 - (\omega_\beta - 1)I.$$

7. *The EWA forward* SOR.

$$M_{E\omega_\beta} = (1/\omega_\beta)(I - \omega_\beta LD_E^{-1})D_E(I - D_E^{-1}U),$$

$$N_{E\omega_\beta} = (1/\omega_\beta)[\omega_\beta N_E - (\omega_\beta - 1)D_E(I - D_E^{-1}U)],$$

$$\mathscr{E}_{\omega_\beta} = M_{E\omega_\beta}^{-1}N_{E\omega_\beta}.$$

8. *The AGA forward* SOR.

$$M_{A\omega_\beta} = (1/\omega_\beta)[I - \omega_\beta(L + H_A)D_A^{-1}]D_A[I - D_A^{-1}(U + Q_A)],$$

$$N_{A\omega_\beta} = (1/\omega_\beta)\{\omega_\beta N_A - (\omega_\beta - 1)D_A[I - D_A^{-1}(U + Q)]\},$$

$$\mathscr{A}_{\omega_\beta} = M_{A\omega_\beta}^{-1} N_{A\omega_\beta}.$$

Introducing a modification similar only to the backward substitution sweep, we obtain

$$(22)\qquad \begin{aligned} \beta^{(t+1)} &= [(L + H)D^{-1}\beta^{(t+1)} + N\phi^{(t)} + c], \\ \phi^{(t+1)} &= \omega_\phi D^{-1}[(U + Q)\phi^{(t+1)} + \beta^{(t+1)}] - (\omega_\phi - 1)\phi^{(t)}, \qquad t \geqq 0. \end{aligned}$$

This method is called the *backward successive overrelaxation two-sweep iterative method* (*backward* SOR) and ω_ϕ, the *relaxation factor*. The elimination of the vector β allows us to reduce (22) to the following equation:

$$(23)\qquad M_{\omega_\phi}\phi^{(t+1)} = N_{\omega_\phi}\phi^{(t)} + c,$$

where

$$(24)\qquad M_{\omega_\phi} = (1/\omega_\phi)[I - (L + H)D^{-1}]D[I - \omega_\phi D^{-1}(U + Q)],$$

$$(25)\qquad N_{\omega_\phi} = (1/\omega_\phi)\{\omega_\phi N - (\omega_\phi - 1)[I - (L + H)D^{-1}]D\},$$

and

$$(26)\qquad \begin{aligned} \mathscr{F}_{\omega_\phi} &= [I - \omega_\phi D^{-1}(U + Q)]^{-1}D^{-1}[I - (L + H)D^{-1}]^{-1} \\ &\quad \times \{\omega_\phi N - (\omega_\phi - 1)[I - (L + H)D^{-1}]D\}. \end{aligned}$$

Consequently, we have the following methods.

9. *The Jacobi backward* SOR.

$$M_{J\omega_\phi} = (1/\omega_\phi)K,$$

$$N_{J\omega_\phi} = (1/\omega_\phi)[\omega_\phi(L + U) - (\omega_\phi - 1)K],$$

$$\mathscr{B}_{\omega_\phi} = M_{J\omega_\phi}^{-1}N_{J\omega_\phi} = \omega_\phi\mathscr{B}_1 - (\omega_\phi - 1)I.$$

10. *The Gauss-Seidel backward* SOR.

$$M_{G\omega_\phi} = (1/\omega_\phi)K[I - \omega_\phi K^{-1}U],$$

$$N_{G\omega_\phi} = (1/\omega_\phi)[\omega_\phi L - (\omega_\phi - 1)K],$$

$$\mathscr{L}_{\omega_\phi} = M_{G\omega_\phi}^{-1}N_{\omega_\phi} = [I - \omega_\phi K^{-1}U]^{-1}K^{-1}[\omega_\phi L - (\omega_\phi - 1)K].$$

11. *The EWA backward* SOR.

$$M_{E\omega_\phi} = (1/\omega_\phi)[I - LD_E^{-1}]D_E[I - \omega_\phi D_E^{-1}U],$$

$$N_{E\omega_\phi} = (1/\omega_\phi)\{\omega_\phi N_E - (\omega_\phi - 1)[I - LD_E^{-1}]D_E\},$$

$$\mathscr{E}_{\omega_\phi} = M_{E\omega_\phi}^{-1}N_{E\omega_\phi}.$$

12. *The AGA backward* SOR.

$$M_{A\omega_\phi} = (1/\omega_\phi)[I - (L + H_A)D_A^{-1}]D_A[I - \omega_\phi D_A^{-1}(U + Q_A)],$$

$$N_{A\omega_\phi} = (1/\omega_\phi)\{\omega_\phi N_A - (\omega_\phi - 1)[I - (L + H_A)D_A^{-1}]D_A\},$$

$$\mathscr{A}_{\omega_\phi} = M_{A\omega_\phi}^{-1}N_{A\omega_\phi}.$$

As can be seen, the Jacobi forward SOR is identical to both the Jacobi backward SOR, and the JOR method of [11]. The Gauss–Seidel backward SOR is exactly equivalent to the standard method known as *the point* SOR *method* [7], [11], however, it is defined in reverse order, as is the Gauss–Seidel method.

3. Double splitting iterative schemes. Expressing the matrix A in the form

$$A = P - R + S, \tag{27}$$

called the *double splitting* of A, where P is a nonsingular matrix, we obtain the following iterative scheme spanned on three successive iterates

$$P\phi^{(t+1)} = R\phi^{(t)} - S\phi^{(t-1)} + c, \qquad t > 0, \tag{28}$$

or, equivalently,

$$\phi^{(t+1)} = P^{-1}R\phi^{(t)} - P^{-1}S\phi^{(t-1)} + P^{-1}c, \qquad t > 0. \tag{29}$$

In the convergence analysis of this scheme, the same approach can be used as in iterative methods based on the single splitting of A ($A = M - N$, where M is a nonsingular matrix). Following the idea of Golub and Varga [8], we can write (29) in the following equivalent form:

$$\begin{bmatrix} \phi^{(t+1)} \\ \phi^{(t)} \end{bmatrix} = \begin{bmatrix} P^{-1}R, & -P^{-1}S \\ I, & 0 \end{bmatrix} \begin{bmatrix} \phi^{(t)} \\ \phi^{(t-1)} \end{bmatrix} + \begin{bmatrix} P^{-1}c \\ 0 \end{bmatrix}. \tag{30}$$

Denoting by

$$\psi^{(t+1)} = \begin{bmatrix} \phi^{(t+1)} \\ \phi^{(t)} \end{bmatrix}, \quad \psi^{(t)} = \begin{bmatrix} \phi^{(t)} \\ \phi^{(t-1)} \end{bmatrix}, \quad v = \begin{bmatrix} P^{-1}c \\ 0 \end{bmatrix} \tag{31}$$

and

$$\mathscr{W} = \begin{bmatrix} P^{-1}R, & -P^{-1}S \\ I, & 0 \end{bmatrix}, \tag{32}$$

we obtain

$$\psi^{(t+1)} = \mathscr{W}\psi^{(t)} + v, \qquad t > 0. \tag{33}$$

Thus the necessary and sufficient condition, which ensures that the iterative method given by (29) tends to the unique solution vector of (1) for all vectors $\phi^{(0)}$ and $\phi^{(1)}$, is that the spectral radius of $\mathscr{W}$, $\rho(\mathscr{W})$, is less than unity. It is interesting to note that when S is a nonsingular matrix, the matrix $\mathscr{W}$ is also nonsingular and

$$\mathscr{W}^{-1} = \begin{bmatrix} 0, & I \\ -S^{-1}P, & S^{-1}R \end{bmatrix}. \tag{34}$$

Applying the SOR procedure simultaneously to both forward and backward sweep equations (15), we obtain

$$\begin{aligned} &\mathscr{B}^{(t+1)} = \omega_\beta[(L + H)D^{-1}\beta^{(t+1)} + N\phi^{(t)} + c] - (\omega_\beta - 1)\beta^{(t)}, \\ &\phi^{(t+1)} = \omega_\phi D^{-1}[(U + Q)\phi^{(t+1)} + \beta^{(t+1)}] - (\omega_\phi - 1)\phi^{(t)}, \qquad t \geqq 0. \end{aligned} \tag{35}$$

The elimination of the vector β allows us to reduce the above equations to (29), in which

$$P = \frac{1}{\omega_\beta \omega_\phi} [I - \omega_\beta(L + H)D^{-1}]D[I - \omega_\phi D^{-1}(U + Q)], \tag{36}$$

$$\begin{aligned} R = \frac{1}{\omega_\beta \omega_\phi} \{ \omega_\beta \omega_\phi N - (\omega_\phi - 1)[I - \omega_\beta(L + H)D^{-1}]D \\ - (\omega_\beta - 1)D[I - \omega_\phi D^{-1}(U + Q)]\}, \end{aligned} \tag{37}$$

$$S = \frac{(\omega_\beta - 1)(\omega_\phi - 1)}{\omega_\beta \omega_\phi} D, \quad \omega_\beta \neq 0, \quad \omega_\phi \neq 0. \tag{38}$$

This method is called the *double successive overrelaxation two-sweep iterative method* or, for brevity, the *double* SOR *method*. It is evident that the method reduces to the single SOR methods—either the forward type when $\omega_\phi = 1$, or the backward type when $\omega_\beta = 1$.

Referring to the former definitions of the matrices H and Q, we have the following methods.

1. *The Jacobi double* SOR.

$$P_J = \frac{1}{\omega_\beta \omega_\phi} K,$$

$$R_J = \frac{1}{\omega_\beta \omega_\phi} [\omega_\beta \omega_\phi (L + U) - (\omega_\beta - 1)K - (\omega_\phi - 1)K],$$

$$S_J = \frac{(\omega_\beta - 1)(\omega_\phi - 1)}{\omega_\beta \omega_\phi} K.$$

2. *The Gauss–Seidel double* SOR.

$$P_G = \frac{1}{\omega_\beta \omega_\phi} K[I - \omega_\phi K^{-1} U],$$

$$R_G = \frac{1}{\omega_\beta \omega_\phi} [\omega_\beta \omega_\phi L - (\omega_\beta - 1)K(I - \omega_\phi K^{-1} U) - (\omega_\phi - 1)K],$$

$$S_G = \frac{(\omega_\beta - 1)(\omega_\phi - 1)}{\omega_\beta \omega_\phi} K.$$

3. *The EWA double* SOR.

$$P_E = \frac{1}{\omega_\beta \omega_\phi} [I - \omega_\beta L D_E^{-1}] D_E [I - \omega_\phi D_E^{-1} U]$$

$$R_E = \frac{1}{\omega_\beta \omega_\phi} [\omega_\beta \omega_\phi N_E - (\omega_\beta - 1) D_E (I - \omega_\phi D_E^{-1} U) - (\omega_\phi - 1)(I - \omega_\beta L D_E^{-1}) D_E],$$

$$S_E = \frac{(\omega_\beta - 1)(\omega_\phi - 1)}{\omega_\beta \omega_\phi} D_E.$$

4. *The AGA double* SOR.

$$P_A = \frac{1}{\omega_\beta \omega_\phi}[I - \omega_\beta(L + H_A)D_A^{-1}]D_A[I - \omega_\phi D_A^{-1}(U + Q_A)],$$

$$R_A = \frac{1}{\omega_\beta \omega_\phi}\{\omega_\beta \omega_\phi N_A - (\omega_\beta - 1)D_A[I - \omega_\phi D_A^{-1}(U + Q_A)]$$

$$- (\omega_\phi - 1)[I - \omega_\beta(L + H_A)D_\mathrm{A}^{-1}]D_A\},$$

$$S_A = \frac{(\omega_\beta - 1)(\omega_\phi - 1)}{\omega_\beta \omega_\phi}D_A.$$

4. Determination of the optimum relaxation factors. Let us examine now the dependence of the iteration matrix eigenvalues on values ω, in order to determine the values of optimum relaxation factors which minimize the spectral radius of the iteration matrix. Since the single (both forward and backward) SOR methods can be considered as special cases of the double SOR method, let us first discuss this method.

4.1. Double SOR two-sweep iterative methods. Both submatrices P and S are nonsingular for all $\omega_\beta \neq 1$ and $\omega_\phi \neq 1$ (and also $\omega_\beta \neq 0$ and $\omega_\phi \neq 0$), but this ensures that the matrix $\mathcal{W}$ is also nonsingular and there exist nonzero vectors $z^T = [x, y]^T$ with corresponding eigenvalues $\nu \neq 0$. Thus,

$$\begin{bmatrix} P^{-1}R, & -P^{-1}S \\ I, & 0 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \nu \begin{bmatrix} x \\ y \end{bmatrix} \tag{39}$$

or

$$P^{-1}Rx - P^{-1}Sy = \nu x,$$

$$x = \nu y,$$

and hence

$$P^{-1}[R - (1/\nu)S]x = \nu x. \tag{40}$$

After some algebra, the above equations can be transformed to the following equation

$$\begin{aligned} D^{-1}[(1 + \theta_\beta\alpha)(L + H) + (1 + \theta_\phi\alpha)(U + Q) - H - Q \\ + \alpha N - \operatorname{diag}\{(L + H)D^{-1}(U + Q)\}]x = \zeta x, \end{aligned} \tag{41}$$

where

$$\alpha = (1 - \nu)/\nu, \qquad \theta_\beta = (\omega_\beta - 1)/\omega_\beta,$$

$$\theta_\phi = (\omega_\phi - 1)/\omega_\phi, \qquad \zeta = (1 + \theta_\beta\alpha)(1 + \theta_\phi\alpha).$$

Now we will examine, in particular iterative schemes, the behaviour of α as a function of θ which corresponds implicitly to the behaviour of $\nu = 1/(1 + \alpha)$ as a function of $\omega = 1/(1 - \theta)$.

A. *The Jacobi double* SOR. Referring to (41) and substituting $H = -L$ and $Q = -U$, we obtain

$$(1 + \alpha)K^{-1}(L + U)x = (1 + \theta_\beta\alpha)(1 + \theta_\phi\alpha)x \tag{42}$$

and

$$(1 + \theta_\beta\alpha)(1 + \theta_\phi\alpha) = \mu(1 + \alpha), \tag{43}$$

where μ are eigenvalues of $\mathscr{B}_1 = K^{-1}(L + U)$.

When $\omega_\beta = \omega_\phi = \omega$ $(\theta_\beta = \theta_\phi = \theta)$, we get

$$\theta^2\alpha^2 - (\mu - 2\theta)\alpha + 1 - \mu = 0, \tag{44}$$

or equivalently,

$$\nu^2 - [\mu\omega^2 - 2(\omega - 1)]\nu + (\omega - 1)^2 = 0. \tag{45}$$

Each eigenvalue μ corresponds two eigenvalues of the iteration matrix $\mathscr{W}_J$

$$\nu_1, \nu_2 = \mu\omega^2/2 - (\omega - 1) \pm \omega[\mu[\mu\omega^2/4 - (\omega - 1)]]^{1/2}, \tag{46}$$

and their values coincide when

$$\mu\omega^2/4 - (\omega - 1) = 0, \tag{47}$$

that is, at the values equal to $\bar{\omega} - 1$ where

$$\bar{\omega} = 2/[1 \pm (1 - \mu)^{1/2}]. \tag{48}$$

When $\mu = 0$, $\nu_1 = \nu_2 = -(\omega - 1)$.

B. *The Gauss–Seidel double* SOR. Substituting $H = -L$ and $Q = 0$ into (41) gives us

$$K^{-1}[(1 + \alpha)L + (1 + \theta_\beta\alpha)U]x = (1 + \theta_\beta\alpha)(1 + \theta_\phi\alpha)x. \tag{49}$$

In the analysis of this method, some theoretical results for matrix problems in which matrices $A = K - L - U$ have certain special properties (namely, they are *p-cyclic consistently ordered* [7]) will be derived. In this case the eigenvalues of the matrix

$$\mathscr{B}(\gamma) = K^{-1}(\gamma L + \gamma^{-(p-1)}U) \tag{50}$$

derived from the Jacobi matrix $\mathscr{B}_1 = K^{-1}(L + U)$ are independent of the parameter $\gamma \neq 0$. Then the matrix $\mathscr{B}_1$ is consistently ordered as well. It is easy to verify that

$$\begin{aligned} \mathscr{B}(\kappa, \sigma) &= K^{-1}[\kappa L + \sigma U] \\ &= (\kappa^{p-1}\sigma)^{1/p}K^{-1}[(\kappa/\sigma)^{1/p}L + [(\kappa/\sigma)^{1/p}]^{-1}U] \end{aligned} \tag{51}$$

for $\kappa \neq 0$, $\sigma \neq 0$, and $p \neq 0$. Thus, if μ and τ are eigenvalues of $\mathscr{B}_1$ and $\mathscr{B}(\kappa, \sigma)$, respectively, then the following relation holds:

$$\tau = \mu(\kappa^{p-1}\sigma)^{1/p}. \tag{52}$$

Referring to (49) and (51), and taking into account the relation (52), we obtain

$$(1 + \theta_\beta\alpha)^{p-1}(1 + \theta_\phi\alpha)^p = \mu^p(1 + \alpha)^{p-1}. \tag{53}$$

Let us restrict the discussion here to the case when $p = 2$. Such 2-cyclic matrices appear in the finite difference approximations to the multidimensional neutron diffusion equations in the rectangular geometry. Hereafter, it is assumed that $\omega_\beta = \omega_\phi = \omega$ $(\theta_\beta = \theta_\phi = \theta)$ and for $p = 2$, we obtain

$$\theta^3\alpha^3 + 3\theta^2\alpha^2 - (\lambda - 3\theta)\alpha + 1 - \lambda = 0, \tag{54}$$

where $\lambda = \mu^2$ is the eigenvalue of the Gauss–Seidel matrix $\mathscr{L}_1$.

It should be mentioned that the substitution of $\alpha = (1 - \nu)/\nu$ and $\theta = (\omega - 1)/\omega$ into the above equation representing sixth-order curve indeed provides the explicit equation for ν and ω, but its analysis is very complicated because multiterm coefficients appear in the equation.

In the case when $0 < \lambda < 1$, the two positive roots ν_1 and ν_2 are minimized when

$$\bar{\omega} = 3/[1 - 2\cos(\varphi/3 + 120^0)], \tag{55}$$

where $\cos\varphi = 2\lambda - 1$ and

$$\nu_1 = \nu_2 = \bar{\nu} = 2(\bar{\omega} - 1) \quad \text{and} \quad \nu_3 = -(\bar{\omega} - 1)/4. \tag{56}$$

For $\omega = 0$, $\nu_1 = \nu_2 = \nu_3 = 1$,

$0 < \omega < 1$, $\lambda < \nu_1 < 1$, ν_2 and ν_3 are complex where $|\nu_2| < \nu_1$,

$\omega = 1$, $\nu_1 = \lambda$ and $\nu_2 = \nu_3 = 0$,

$1 < \omega < \bar{\omega}$, $0 < \nu_2 < \nu_1 < \lambda$ and $0 > \nu_3 > -(\omega - 1)/4$,

$\omega = \bar{\omega}$, $\nu_1 = \nu_2 = 2(\bar{\omega} - 1)$ and $\nu_3 = -(\bar{\omega} - 1)/4$,

$\bar{\omega} < \omega$, ν_1 and ν_2 are complex where $|\nu_1| > \omega - 1$

and $-(\omega - 1)/4 > \nu_3 > -1/4$.

When $\lambda = 0$, $\nu_1 = \nu_2 = \nu_3 = -(\omega - 1)$.

Figure 1 shows the graph of the roots ν versus ω for $\lambda = \mu^2 = 0.9$ in the Gauss–Seidel double SOR method for which the corresponding curves are denoted by the letter D. Curves drawn by continuous lines represent real and positive values of ν and dashed lines represent the modulus of complex values of ν. The dotted-dashed lines correspond to the modulus of real and negative values of ν. The points corresponding to the values of $\bar{\nu}$ are marked by the digit 1.

Usually the term $\theta^3\alpha^3$ is a few orders of magnitude smaller than the remaining terms in (54). By neglecting this term and solving the following reduced equation

$$3\theta^2\alpha^2 - (\lambda - 3\theta)\alpha + 1 - \lambda = 0, \tag{57}$$

we obtain

$$\bar{\omega}_a = 1/(1 - \bar{\theta}_a), \tag{58}$$

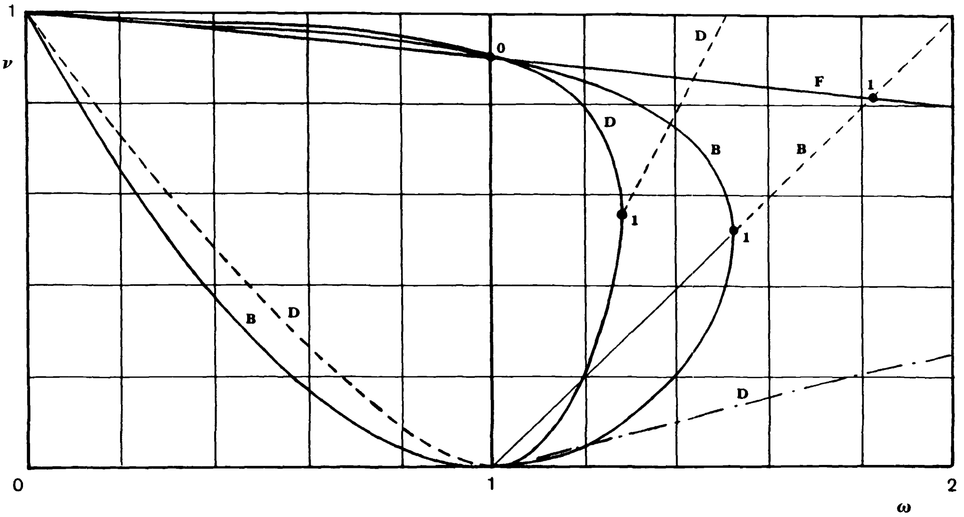


FIG. 1. *RECSOR*-II.

where

$$\bar{\theta}_a = \lambda/[3 + 2(3 - 3\lambda)^{1/2}]$$

and

(59) $$\bar{\nu}_a = (\bar{\omega}_a - 1)/[\bar{\omega}_a[1 + ((1 - \lambda)/3)^{1/2}] - 1].$$

As can be seen from Table 1, the values of $\bar{\omega}_a$ and $\bar{\nu}_a$ calculated from (58) and (59) provide quite a good approximation to the exact values of $\bar{\omega}$ and $\bar{\nu}$ obtained from (55) and (56).

Finally it should be mentioned that for the Jacobi double SOR method, when $p = 2$, the value of ω which minimizes the spectral radius of $\mathscr{W}_J$ is $\bar{\omega} = 1$.

C. *The AGA* (*EWA*) *double* SOR *method*. In applying this method, representing a wide class of algorithms (including its special case, the EWA method, and defined by $H = Q = 0$), to the same matrix problems, we observed that the behaviour of roots ν versus ω was similar to the Gauss–Seidel double SOR method. However, for the optimum values $\bar{\omega}$ for which the values of ν_1 and ν_2 coincide ($\nu_1 = \nu_2 = \bar{\nu}$), new eigenvalues ν_i derived from the negative (or complex) eigenvalues of the matrix $\mathscr{A}_1$ or $\mathscr{E}_1$ occur at the absolute value, which is greater than $\bar{\nu}$. In graphic presentation, the cut of the curve representing ν_1 and ν_2 by the curve corresponding to the absolute value of ν_i is for $\omega = \omega_c < \bar{\omega}$. This proves that the spectral radius of $\mathscr{W}_A$ is greater for $\omega = \bar{\omega}$ than that for $\omega = \omega_c$. However, as is observed in practice, for $\omega = \bar{\omega}$ the oscillation character of the convergence is occurring and the number of iterations for a given accuracy of the solution is usually smaller than the number of iterations required for $\omega = \omega_c$.

When the $n \times n$ matrix A is 2-cyclic and consistently ordered, the matrix $\mathscr{B}_1$ has n simple eigenvalues, namely, $\pm\mu_1, \pm\mu_2, \ldots, \pm\mu_{n/2}$ (if n is odd, there is an extra zero eigenvalue); the matrix $\mathscr{L}_1$ has $n/2$ eigenvalues $\mu_1^2, \mu_2^2, \ldots$, and all the other eigenvalues are zeros. Changing over to the AGA (EWA) factorizations, negative and/or complex eigenvalues appear in the eigenvalue spectrum of matrices $\mathscr{A}_1$ and $\mathscr{E}_1$.

It is interesting to note that in the AGA double SOR method, curves representing the behaviour of ν_1 and ν_2 versus ω are perfectly approximated in the range $0 < \omega < \bar{\omega}$ by the equation

(60) $$A\theta^3\alpha^3 + B\theta^2\alpha^2 - (\lambda - C\theta)\alpha + 1 - \lambda = 0,$$

in which coefficients A, B, and C can be calculated from three equations given by three values of ν computed for three different values of ω, for instance, by power method and using the substitutions $\alpha = (1 - \nu)/\nu$ and $\theta = (\omega - 1)/\omega$.

In practice, the same computational procedure as in the Gauss–Seidel double SOR method can be based on the solution of the reduced form of (60), that is,

(61) $$\bar{B}\theta^2\alpha^2 - (\lambda - \bar{C}\theta)\alpha + 1 - \lambda = 0,$$

TABLE 1

$\lambda = \mu^2$	$\bar{\omega}$	$\bar{\nu}$	$\bar{\omega}_a$	$\bar{\nu}_a$	$\bar{\omega}_b$	$\bar{\nu}_b$
1.	1.5	1.	1.5	1.	2.	1.
0.999	1.4733	0.9466	1.4733	0.9462	1.9387	0.9387
0.99	1.4196	0.8392	1.4201	0.8367	1.8182	0.8182
0.9	1.2789	0.5578	1.2817	0.5462	1.5195	0.5195
0.5	1.0981	0.1962	1.1010	0.1835	1.1716	0.1716
0.1	1.0155	0.0310	1.0162	0.0282	1.0263	0.0263
0.	1.	0.	1.	0.	1.	0.

which provides the estimation of $\bar{\omega}$ with the accuracy of four significant figures. This is completely sufficient because changes of the fourth significant digit in ω do not change the number of iterations for a given accuracy of the solution. The computer time of this procedure is comparable (very often shorter in large problems) with the time required only for the calculation of the spectral radius of $\mathscr{L}_1$. Usually the *dominance ratio* defined as

$$\sigma = |\lambda_1|/|\lambda_2|, \tag{62}$$

where the eigenvalues have the following order $|\lambda_1| > |\lambda_2| > |\lambda_3|, \ldots,$ is significantly smaller for the matrix $\mathscr{A}_1$ than in the case of the matrix $\mathscr{L}_1$. The behaviour of ν versus ω is illustrated in Figs. 2, 3, and 4 in § 5.

4.2. Single SOR two-sweep iterative methods. As mentioned, these methods can be considered as special cases of the double SOR methods in which either $\omega_\beta \neq 1$ and $\omega_\phi = 1$ (forward type) or $\omega_\beta = 1$ and $\omega_\phi \neq 1$ (backward type).

A. *The Jacobi single* SOR *method*. Assuming either $\omega_\beta = \omega$ and $\omega_\phi = 1$ or $\omega_\beta = 1$ and $\omega_\phi = \omega$ in (43), we obtain for both cases the same equation,

$$1 + \theta\alpha = \mu(1 + \alpha), \tag{63}$$

and by substituting $\alpha = (1 - \nu)/\nu$ and $\theta = (\omega - 1)/\omega$,

$$\nu = \omega\mu - (\omega - 1). \tag{64}$$

When $\mathscr{B}_1$ is a 2-cyclic and consistently ordered matrix, the value of $\omega = 1$ minimalizes the spectral radius of $\mathscr{B}_\omega$.

B. *The Gauss–Seidel single* SOR *method*. In the forward type, in which $\omega_\beta = \omega$ and $\omega_\phi = 1$, we obtain from the matrix $\mathscr{L}_{\omega_\beta}$ the relation

$$\nu = \omega\lambda - (\omega - 1), \tag{65}$$

where λ is the eigenvalue of $\mathscr{L}_1$. For the 2-cyclic consistently ordered case the spectral radius of $\mathscr{L}_\omega$ is minimized for

$$\omega = 2/(2 - \lambda). \tag{66}$$

In the backward type, the substitution of $\omega_\beta = 1$ and $\omega_\phi = \omega$, and $p = 2$ in (53), gives

$$(1 + \theta\alpha)^2 = \mu^2(1 + \alpha) \tag{67}$$

or

$$\theta^2\alpha^2 - (\lambda - 2\theta)\alpha + 1 - \lambda = 0. \tag{68}$$

The solution of this equation gives the well-known formula

$$\bar{\omega}_b = 2/[1 + (1 - \lambda)^{1/2}] \tag{69}$$

for the point SOR method [7], where $\lambda = \mu^2$ and $\bar{\nu}_b = \rho(\mathscr{L}_{\omega_b}) = \bar{\omega}_b - 1$.

The behaviour of roots ν versus ω is plotted in Fig. 1 for $\lambda = 0.9$, where the curve denoted by B corresponds to the Gauss–Seidel backward SOR (the point SOR), the curve denoted by F to the Gauss–Seidel forward SOR, and the curve denoted by D to the Gauss–Seidel double SOR method. The points denoted by digit 1 correspond to optimal values of ν. The algorithm of this method, when applied to the discrete solution of two-dimensional elliptic problems in rectangular geometry, is called RECSOR-II.

The values of $\bar{\omega}_b$ and $\bar{\nu}_b$ for different values of $\lambda = \mu^2$ are given in the table. However in spite of the fact that

$$\rho(\mathscr{L}_{\bar{\omega}}) < \rho(\mathscr{W}_G),$$

both methods are almost equivalent from the viewpoint of convergence properties; the ratio of $R_\infty(\mathscr{L}_{\bar{\omega}})$ to $R_\infty(\mathscr{W}_G)$ changes in the range $1.15 \div 1.05$ for the values of λ given in the table.

C. *The AGA* (*EWA*) *single* SOR *methods*. In algorithms of this method in which the nonzero diagonals of matrices N_A are located symmetrically with respect to the main diagonal, both the forward and backward SOR processes give almost the same values of $\bar{\omega}$ and ω_c, as well as the values of $\bar{\nu}$ and ν_c. The value of ν_c corresponding to the minimum spectral radius is much greater than $\bar{\nu}$ and, for this reason, these methods are less recommended than the point SOR. This effect is illustrated in the next section.

For matrices N_A with the unsymmetrical locations of nonzero diagonals with respect to the main diagonal, the rate of convergence in the forward AGA method can be greater than in the case of the backward AGA method. This effect is contrary to the behaviour of the Gauss–Seidel forward and backward SOR methods.

5. Model problem analysis. Let us consider the solution of the two-dimensional Dirichlet problem for the Laplace operator

$$-\nabla^2\Phi(x, v) = f(x, v), \qquad (x, v) \in R. \tag{70}$$

The problem is solved in the hexagonal geometry oriented in the oblique coordinate system (x, v) where a uniform triangular mesh is imposed on the unit 120-degree parallelogram consisting of $s + 1$ oblique columns and t rows [3]. The condition of reflection symmetry is assumed on the left, top, and bottom boundaries, and $\Phi = 0$ on the right boundary of this parallelogram.

The seven-point finite difference approximation leads to the linear equation system in which an $n \times n$ matrix A has seven nonzero diagonals where $n = s \times t$. We assume zero as the index for the main diagonal, and $1, 2, \ldots$ for the indices of the successive diagonals in the upper part of A, and $-1, -2, \ldots$ for the indices of the successive diagonals in the lower part of A; the matrix A in the above problem can then be described by the following nonzero diagonal indices, $-s - 1$, $-s$, -1, 0, 1, s, $s + 1$. Such matrices are not p-cyclic and not consistently ordered.

Let us restrict our attention to the spectral radius analysis of the iteration matrices in the above problem for the point SOR method, called HEXSOR-II here, and to two algorithms of the AGA method, called the HEXAGA-II-A [3] and HEXAGA-II-A1. In HEXAGA-II-A, the nonzero diagonal indices for the matrix H_A are $-s$ and -1; for the matrix Q_A, they are 1 and s; and for the matrix N_A, they are $-s + 1$ and $s - 1$. In the case of HEXAGA-II-A1, the nonzero diagonal indices are $-s$, $-s + 1$, and -1 for H_A; 1 and s for Q_A; and $-s + 2$, 2, and $s - 1$ for N_A.

The behaviour of eigenvalues ν versus ω for the iteration matrices in HEXSOR-II, HEXAGA-II-A, and HEXAGA-II-A1 is shown in Figs. 2, 3, and 4, respectively, for the case when $s = 6$ and $t = 3$.

The EISRG2 subroutine from the EISPACK library was used for the calculation of the eigenvalue spectrum of iteration matrices on the PC-AT computer. The explanation of symbols and curves on these figures is the same as for Fig. 1. The points denoted by digit 2 correspond to the value of ω_c for which the spectral radius has the minimum value. In Fig. 2 the curve denoted by R represents the case when the matrix A is 2-cyclic and consistently ordered (but with the same value of spectral radius for $\omega = 1$). What

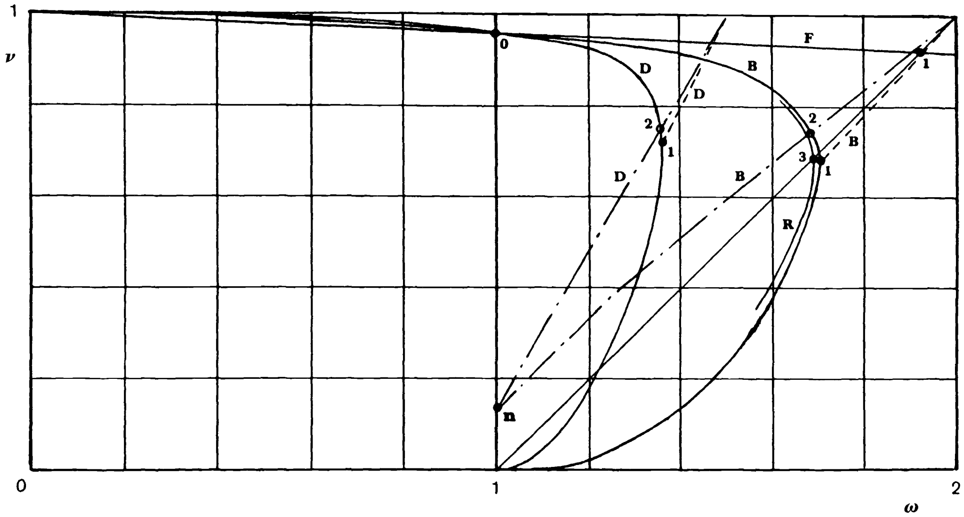


FIG. 2. *HEXSOR*-II.

takes place in the rectangular geometry and the minimum value of spectral radius is at the point denoted by digit 3. This illustration allows us to conclude that the deviation of the convergence behaviour of the point SOR method (HEXSOR-II) in the hexagonal geometry is very small compared to that in the rectangular geometry (RECSOR-II).

As can be seen from this graph representation, the minimum values of spectral radii are obtained for HEXAGA-II-A and HEXAGA-II-A1 when the double SOR process is applied. In calculations, it was assumed that in the double SOR $\omega_\beta = \omega_\phi = \omega$.

It should be noted that the behaviour of spectral radii as a function of ω shown in Figs. 2, 3, and 4 is representative for multidimensional neutron diffusion equation problems solved by these algorithms. The results of benchmark problem calculations by the HEXAGA programs in comparsion to other programs are demonstrated in [9].

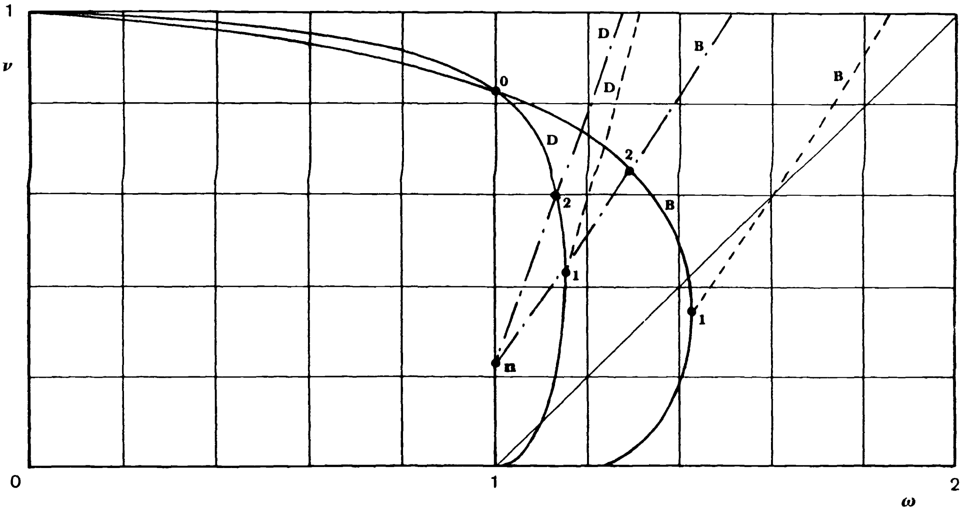


FIG. 3. *HEXAGA*-II-*A*.

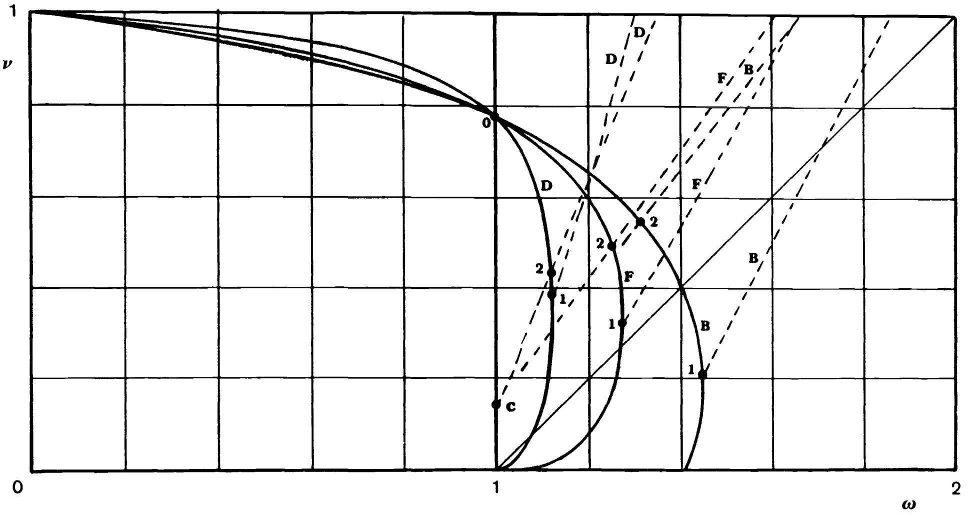


FIG. 4. *HEXAGA*-II-*A*1.

Acknowledgments. The author would like to acknowledge the referee for his helpful comments and Dr. K. Masłowski for his contribution to numerical results.

REFERENCES

[1] Z. I. Woźnicki. *Two-sweep iterative methods for solving large linear systems and their applications to the numerical solution of multigroup, multidimensional neutron diffusion equations*. Ph.D. thesis, Rep. No 1447-CYFRONET-PM-A, Inst. of Nuclear Research, Swierk-Otwock, Poland, 1973.

[2] ———, *AGA two-sweep iterative methods and their application in critical reactor calculations*, Nucleonika, 23 (1978), pp. 941–967.

[3] ———, *HEXAGA*-II-120, -60, -30 *two-dimensional multigroup neutron diffusion programmes for a uniform triangular mesh with arbitrary group scattering*, Rep. KfK 2789, Kernforschunszentrum Karlsruhe, Karlsruhe, Germany, 1979.

[4] ———, *HEXAGA*-III-120, -30 *three-dimensional multigroup neutron diffusion programmes for a uniform triangular mesh with arbitrary group scattering*, Rep. KfK 3572, Kernforschungszentrum Karlsruhe, Karlsruhe, Germany, 1983.

[5] R. Koch and Z. I. Woźnicki, *The classification of partial factorization iterative methods used for solving linear equation systems*, in Proc. Internat. Topical Meeting on Advances in Reactor Physics, Mathematics, and Computations. Paris, France, Apr. 1987.

[6] Z. I. Woźnicki, *AGA two-sweep iterative methods and their application for the solution of linear equation systems*, in Proc. Internat. Conf. on Linear Algebra and Applications, Valencia, Spain, Sept. 1987; Linear Algebra Appl., 121 (1989), pp. 702–710.

[7] R. S. Varga, *Matrix Iterative Analysis* Prentice-Hall, Englewood Cliffs, NJ, 1962.

[8] G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods*. I. Numer. Math., 3 (1961), pp. 147–168.

[9] Z. I. Woźnicki, *Two- and three-dimensional benchmark calculations for triangular geometry by means of HEXAGA programmes*, in Proc. Internat. Meeting on Advances in Nuclear Engineering Computational Methods, Knoxville, TN, April 1985, American Nuclear Society, Inc., LaGrange Park, IL.

[10] ———, *EWA*-II *two-dimensional, two-group diffusion fast code*, Kernenergie, 10 (1971), pp. 323–328.

[11] D. M. Young, *Iterative solution of large linear systems*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 74-81, January 1993

006

ON A CONJECTURE OF PIERCE FOR PERMANENTS OF SINGULAR CORRELATION MATRICES*

C. L. FRENZEN† AND I. FISCHER†

Abstract. Let I_n be the n-by-n identity matrix, let J_n be the n-by-n matrix, all of whose entries equal 1, and let $Y_n = n/(n-1)I_n - 1/(n-1)J_n$. Y_n is then a singular correlation matrix, that is, Y_n is a singular positive semidefinite Hermitian matrix all of whose diagonal elements equal 1. Pierce has conjectured that if A is a singular correlation matrix, the minimum value of the permanent of A occurs exactly when A is similar to Y_n by a diagonal unitary matrix. If Pierce's conjecture is true, then the permanent of Y_n (per Y_n) must be a decreasing function of n. This paper proves that per Y_n is strictly decreasing for $n = 2, 3, 4, \ldots$.

Key words. permanent, correlation matrix, asymptotic approximation

AMS(MOS) subject classifications. 15A15, 15A45

1. Introduction. A correlation matrix of order n is an n-by-n positive semidefinite Hermitian matrix all of whose diagonal elements equal 1. Denote the set of all correlation matrices by C_n. When $A, B \in C_n$ and A is similar to B by a diagonal unitary matrix, we write $A \equiv B$. Note that if $A \equiv B$, then per A = per B, where per A and per B represent the permanents of A and B, respectively. In [4] Pierce conjectured that if A is a singular correlation matrix, the minimum value of per A occurs exactly when $A \equiv Y_n$, where

$$Y_n = \frac{n}{n-1} I_n - \frac{1}{n-1} J_n, \tag{1}$$

with I_n being the n-by-n identity matrix and J_n being the n-by-n matrix with all entries equal to 1. Pierce's conjecture was verified for $n = 4$ in [3].

If Pierce's conjecture is true, then

$$\text{per } Y_{n+1} < \text{per } Y_n \qquad (n = 2, 3, 4, \ldots). \tag{2}$$

To see this, form the $(n+1)$-by-$(n+1)$ matrix A from Y_n by letting a_{ij} be the (i, j) element of Y_n for $1 \leqq i \leqq n$ and $1 \leqq j \leqq n$, $a_{n+1,n+1} = 1$. Since $A \in C_{n+1}$, A is singular, and $A \not\equiv Y_n$, the truth of Pierce's conjecture implies per $A >$ per Y_{n+1}. However, by the construction of A, per A = per Y_n, and (2) follows. The first nine values of per Y_n (correct to five places) are given in Table 1.

Pierce's conjecture would be false if (2) failed to hold for some n. The purpose of this paper is to prove that (2) holds for all $n \geqq 2$, thereby removing one means of disproving Pierce's conjecture. We obtain an asymptotic approximation with error bounds for per Y_n from an integral representation by Barrett and Lundquist [1]. Using this approximation, we then show that

$$\text{per } Y_{n+1} - \text{per } Y_n < \frac{e}{8(n+1)^3}(9-n) \qquad (n \geqq 10). \tag{3}$$

Equation (2) then follows from Table 1 and relation (3). We note that several other monotonicity questions have recently been settled by using asymptotic expansions with

* Received by the editors January 30, 1991; accepted for publication (in revised form) July 30, 1991. This work was supported (in part) by the Naval Postgraduate School Research Program.

† Department of Mathematics, Naval Postgraduate School, Monterey, California 93943 (5191p@navpgs.bitnet).

TABLE 1

n	per Y_n
2	2
3	1.5
4	1.48148
5	1.44531
6	1.42848
7	1.41662
8	1.40832
9	1.40209
10	1.39727

error bounds. In [2] and [5] the monotonicity of the Lebesgue constants for Legendre series was established, and in [7] the monotonicity of the points of inflection of the Bessel function J_ν was proved.

2. An integral representation for per Y_n. We begin with the identity

$$\text{per}\,(tI_n + aJ_n) = \sum_{k=0}^{n} \binom{n}{k} (t+a)^k a^{n-k} D_{n-k}, \tag{4}$$

where D_k, the kth derangement number, may be expressed as

$$D_k = k! \sum_{j=0}^{k} \frac{(-1)^j}{j!}. \tag{5}$$

The identity in (4) may be derived by noting that $D_n = \text{per}\,(J_n - I_n)$. By writing $tI_n + aJ_n$ as $(t+a)I_n + a(J_n - I_n)$ we can find the permanent of the latter matrix (all of whose diagonal entries equal $t + a$) by computing the coefficient of $(t+a)^k$ in its permanent for $k = 0, \ldots, n$. Each of the $\binom{n}{k}$ ways of selecting k of the diagonal entries yields a factor of $(t+a)^k$ in the permanent, and the coefficient of this factor is per $[a(J_{n-k} - I_{n-k})] = a^{n-k}D_{n-k}$. Multiplying $(t+a)^k$ by $\binom{n}{k}a^{n-k}D_{n-k}$ and summing on k then gives (4).

Given that

$$\int_0^\infty (1-x)^k e^{-x}\,dx = (-1)^k k! \sum_{j=0}^{k} \frac{(-1)^j}{j!},$$

it follows from (5) that

$$D_k = (-1)^k \int_0^\infty (1-x)^k e^{-x}\,dx. \tag{6}$$

Substituting (6) into (4) then gives

$$\begin{aligned} \text{per}\,(tI_n + aJ_n) &= \int_0^\infty (t+ax)^n e^{-x}\,dx \\ &= n!a^n \sum_{k=0}^{n} \frac{(t/a)^k}{k!}, \end{aligned} \tag{7}$$

where the second equality in (7) comes from applying the binomial theorem to the integrand in (7) and integrating termwise. After we set $t = n/(n-1)$ and $a = -1/(n-1)$, (7) and (1) then yield

$$\begin{aligned}\text{per } Y_n &= \int_0^\infty \left(\frac{n-x}{n-1}\right)^n e^{-x}\,dx \\ &= \frac{(-1)^n n!}{(n-1)^n}\sum_{k=0}^{n-2}\frac{(-1)^k n^k}{k!}.\end{aligned}\tag{8}$$

Equation (8) and its derivation are from Barrett and Lundquist [1].

The entries in Table 1 were computed from the sum in (8). It is not difficult to show from the integral in (8) (see the argument following (10)) that

$$\lim_{n\to\infty} \text{per } Y_n = \frac{e}{2} = 1.3591409142\cdots.$$

We note that the above limit was also observed in [4] by means of a different integral. Although reference was made in [4] to an unpublished proof for asymptotic monotonicity of per Y_n, implying that per $Y_n > e/2$ for all sufficiently large n, the monotonicity result given in the present paper implies that per $Y_n > e/2$ for all $n \geqq 2$.

Using (8) and substituting $x = ny$, we now write

$$\text{per } Y_n = P_1(n) + P_2(n),\tag{9}$$

where

$$\begin{aligned}P_1(n) &= \left(\frac{n}{n-1}\right)^n\left(n\int_0^1 (1-y)^n e^{-ny}\,dy\right), \\ P_2(n) &= (-1)^n\frac{n^{n+1}}{(n-1)^n}\int_1^\infty (y-1)^n e^{-ny}\,dy \\ &= \frac{(-1)^n e^{-n} n!}{(n-1)^n}.\end{aligned}\tag{10}$$

Clearly, $P_2(n) \to 0$ as $n \to \infty$. Now

$$P_1(n) = \left(\frac{n}{n-1}\right)^n \int_0^n \left(1-\frac{x}{n}\right)^n e^{-x}\,dx.$$

The factor in front of the integral above tends to e as $n \to \infty$, while the integral itself tends to

$$\int_0^\infty e^{-2x}\,dx = \frac{1}{2}$$

by the dominated convergence theorem. Thus $P_1(n)$, and therefore per Y_n, tends to $e/2$ as $n \to \infty$.

From [6, Prob. 20, p. 200] we have

$$n! < en^n e^{-n}\sqrt{n} \qquad (n \geqq 2);$$

by applying this to $P_2(n)$ in (10) we conclude that

$$(-1)^n P_2(n) < e\left(\frac{n}{n-1}\right)^n e^{-2n}\sqrt{n}.\tag{11}$$

Now (10) implies that

$$P_2(n+1) = -\frac{(n+1)(n-1)^n}{en^{n+1}} P_2(n),$$

and by using this expression for $P_2(n+1)$ and the bound in (11) we conclude that

$$(12)\quad |P_2(n+1) - P_2(n)| < \frac{e}{2(n+1)^3}[2e^{-2n}\sqrt{n}(n+1)^3]\left[\left(\frac{n}{n-1}\right)^n + \frac{n+1}{ne}\right].$$

Since $2e^{-2n}\sqrt{n}(n+1)^3$, $[n/(n-1)]^n$, and $(n+1)/n$ all decrease with n when $n \geqq 2$, (12) implies that

$$(13)\quad \begin{aligned} |P_2(n+1) - P_2(n)| &< \frac{e}{2(n+1)^3}[2e^{-20}\sqrt{10}(11)^3]\left[\left(\frac{10}{9}\right)^{10} + \frac{11}{10e}\right] \\ &< \frac{e}{2(n+1)^3}(10^{-4}) \qquad (n \geqq 10). \end{aligned}$$

From (9) and (13) it then follows that

$$(14)\quad \text{per } Y_{n+1} - \text{per } Y_n < P_1(n+1) - P_1(n) + \frac{10^{-4}e}{2(n+1)^3} \qquad (n \geqq 10).$$

3. Asymptotic approximation for $P_1(n)$. From (10),

$$(15)\quad P_1(n) = F(n)H(n),$$

where

$$(16)\quad \begin{aligned} F(n) &= \left(\frac{n}{n-1}\right)^n, \\ H(n) &= n\int_0^1 e^{-ny}(1-y)^n\,dy. \end{aligned}$$

We shall now show that the right side of (14) is less than zero for $n \geqq 10$ by deriving an asymptotic approximation with error bounds for $P_1(n)$. $P_1(n)$ decreases with n, but this is not entirely obvious since $F(n)$ decreases with n and $H(n)$ *increases* with n.

From (16),

$$F(n) = \exp\left[-n\ln\left(1-\frac{1}{n}\right)\right] = \exp\left(1 + \frac{1}{2n} + \frac{1}{3n^2} + \cdots\right),$$

so that

$$(17)\quad \frac{F(n)}{e} = \exp\left(\frac{1}{2n} + \frac{1}{3n^2} + \frac{1}{4n^3} + \cdots\right).$$

Equation (17) implies that

$$(18)\quad \begin{aligned} \frac{F(n)}{e} &> \exp\left(\frac{1}{2n} + \frac{1}{3n^2}\right) \\ &> 1 + \left(\frac{1}{2n} + \frac{1}{3n^2}\right) + \frac{1}{2!}\left(\frac{1}{2n} + \frac{1}{3n^2}\right)^2 \\ &> 1 + \frac{1}{2n} + \frac{11}{24n^2} + \frac{1}{6n^3} \qquad (n \geqq 2). \end{aligned}$$

To derive an upper bound for $F(n)/e$, note that

$$\frac{1}{4n^3}+\frac{1}{5n^4}+\cdots<\frac{1}{4n^3}\frac{n}{n-1} \qquad (n \geqq 2),$$

so that

$$\frac{F(n)}{e}<\exp\left(\frac{1}{2n}+\frac{1}{3n^2}+\frac{1}{4n^3}\frac{n}{n-1}\right) \qquad (n \geqq 2). \tag{19}$$

Since the exponent in (19) decreases with n and is less than $\frac{1}{2}$ for $n \geqq 2$, and since

$$e^x<1+x+\frac{x^2}{2}+\frac{x^3}{3} \qquad \left(0<x<\frac{1}{2}\right), \tag{20}$$

letting x be the argument of the exponential in (19) and substituting into (20) then gives

$$\frac{F(n)}{e}<1+\frac{1}{2n}+\frac{11}{24n^2}+\frac{A(n)}{n^3}, \tag{21}$$

where

$$\begin{aligned} A(n)&=\frac{1}{6}+\frac{1}{4}\frac{n}{n-1}+\frac{1}{3}\left(\frac{1}{2}+\frac{1}{3n}+\frac{1}{4n(n-1)}\right)^3+\frac{1}{18n}\\ &+\frac{1}{8(n-1)}+\frac{1}{12n(n-1)}+\frac{1}{32n}\left(\frac{1}{n-1}\right)^2. \end{aligned} \tag{22}$$

Since $A(n)$ decreases with n, (22) implies that

$$A(n)\leqq A(10)<0.54 \qquad (n\geqq 10). \tag{23}$$

Using (23) in (21) then yields

$$\frac{F(n)}{e}<1+\frac{1}{2n}+\frac{11}{24n^2}+\frac{0.54}{n^3} \qquad (n\geqq 10). \tag{24}$$

After (18) and (24) are combined, it then follows that for all $n \geqq 10$

$$1+\frac{1}{2n}+\frac{11}{24n^2}+\frac{1}{6n^3}<\frac{F(n)}{e}<1+\frac{1}{2n}+\frac{11}{24n^2}+\frac{0.54}{n^3}. \tag{25}$$

We shall now derive similar upper and lower bounds for $H(n)$ in (16) by integration by parts. Let

$$h(y)=y-\ln(1-y), \qquad f(y)=\frac{1}{\dfrac{d}{dy}h(y)}=\frac{1-y}{2-y}$$

and

$$g_0(y)=1, \quad g_{j+1}(y)=\frac{d}{dy}[f(y)g_j(y)], \quad j=0,1,2.$$

Integrating by parts then gives

$$\begin{aligned} H(n) &= n\int_0^1 e^{-nh(y)}\,dy = \frac{1}{2} + \int_0^1 e^{-nh(y)} g_1(y)\,dy \\ &= \frac{1}{2} - \frac{1}{8n} + \frac{1}{n}\int_0^1 e^{-nh(y)} g_2(y)\,dy \qquad (26) \\ &= \frac{1}{2} - \frac{1}{8n} - \frac{1}{32n^2} + \frac{1}{n^2}\int_0^1 e^{-nh(y)} g_3(y)\,dy. \end{aligned}$$

Since

$$g_3(y) = \frac{-6}{(2-y)^4} + \frac{20}{(2-y)^5} + \frac{-15}{(2-y)^6},$$

we find after some calculation that

$$\max_{y\in[0,1]} |g_3(y)| = |g_3(1)| = 1. \qquad (27)$$

Using (27) in (26) then yields

$$\left|H(n) - \left(\frac{1}{2} - \frac{1}{8n} - \frac{1}{32n^2}\right)\right| < \frac{1}{n^2}\int_0^1 e^{-nh(y)}\,dy < \frac{1}{n^2}\int_0^1 e^{-2ny}\,dy < \frac{1}{2n^3},$$

so that

$$\frac{1}{2} - \frac{1}{8n} - \frac{1}{32n^2} - \frac{1}{2n^3} < H(n) < \frac{1}{2} - \frac{1}{8n} - \frac{1}{32n^2} + \frac{1}{2n^3}. \qquad (28)$$

Let

$$\begin{aligned} Q(n) &= \left(1 + \frac{1}{2n} + \frac{11}{24n^2}\right)\left(1 - \frac{1}{4n} - \frac{1}{16n^2}\right) \qquad (29) \\ &= 1 + \frac{1}{4n} + \frac{13}{48n^2} - \frac{7}{48n^3} - \frac{11}{384n^4}. \end{aligned}$$

Multiplying the upper bounds in (25) by those in (28) and using (15) then gives

$$\begin{aligned} \frac{P_1(n)}{e/2} &< Q(n) + \frac{0.54}{n^3}\left(1 - \frac{1}{4n} - \frac{1}{16n^2}\right) \qquad (30) \\ &\quad + \frac{1}{n^3}\left(1 + \frac{1}{2n} + \frac{11}{24n^2}\right) + \frac{0.54}{n^6} \qquad (n \geqq 10) \end{aligned}$$

or

$$\frac{P_1(n)}{e/2} < 1 + \frac{1}{4n} + \frac{13}{48n^2} + \frac{B_1(n)}{n^3} \qquad (n \geqq 10), \qquad (31)$$

where

$$0 < B_1(n) < 1.4 + \frac{0.34}{n} + \frac{0.47}{n^2} + \frac{0.54}{n^3} \qquad (n \geqq 10). \qquad (32)$$

Since the right side of (32) decreases with n, evaluating it for $n = 10$ implies that

$$0 < B_1(n) < 1.44 \qquad (n \geqq 10). \tag{33}$$

Using (33) in (31) then yields

$$\frac{P_1(n)}{e/2} < 1 + \frac{1}{4n} + \frac{13}{48n^2} + \frac{1.44}{n^3} \qquad (n \geqq 10). \tag{34}$$

A similar multiplication of the lower bounds in (25) and (28) gives

$$\begin{aligned}\frac{P_1(n)}{e/2} &> Q(n) - \frac{1}{n^3}\left(1 + \frac{1}{2n} + \frac{11}{24n^2}\right) \\ &+ \frac{1}{6n^3}\left(1 - \frac{1}{4n} - \frac{1}{16n^2}\right) - \frac{1}{6n^6} \qquad (n \geqq 10)\end{aligned} \tag{35}$$

or

$$\frac{P_1(n)}{e/2} > 1 + \frac{1}{4n} + \frac{13}{48n^2} - \frac{B_2(n)}{n^3} \qquad (n \geqq 10), \tag{36}$$

where

$$0 < B_2(n) < 1 + \frac{2}{3n} + \frac{1}{2n^2} + \frac{1}{6n^3} \qquad (n \geqq 10). \tag{37}$$

Again the right side of (37) decreases with n, so evaluating it for $n = 10$ implies that

$$0 < B_2(n) < 1.08 \qquad (n \geqq 10). \tag{38}$$

Relations (36) and (38) then yield

$$\frac{P_1(n)}{e/2} > 1 + \frac{1}{4n} + \frac{13}{48n^2} - \frac{1.08}{n^3} \qquad (n \geqq 10). \tag{39}$$

Combining (34) and (39) then gives us the asymptotic approximation for $P_1(n)$ we need:

$$1 + \frac{1}{4n} + \frac{13}{48n^2} - \frac{1.08}{n^3} < \frac{P_1(n)}{e/2} < 1 + \frac{1}{4n} + \frac{13}{48n^2} + \frac{1.44}{n^3} \tag{40}$$

for all $n \geqq 10$.

4. Proof of (3). Relation (40) implies that for $n \geqq 10$

$$\frac{P_1(n+1) - P_1(n)}{e/2} < \frac{-1}{4n(n+1)} - \frac{13}{48}\frac{2n+1}{n^2(n+1)^2} + \frac{1.44}{(n+1)^3} + \frac{1.08}{n^3}. \tag{41}$$

Since

$$\frac{-1}{4n(n+1)} < \frac{-1}{4(n+1)^2}, \qquad \frac{-(2n+1)}{n^2(n+1)^2} < \frac{-2}{(n+1)^3},$$

and

$$\frac{1}{n^3} = \frac{1}{(n+1)^3}\left(1 + \frac{1}{n}\right)^3 \leqq \frac{1}{(n+1)^3}(1.1)^3 < \frac{1.34}{(n+1)^3} \qquad (n \geqq 10),$$

(41) implies that

$$\frac{P_1(n+1)-P_1(n)}{e/2} < \frac{1}{(n+1)^3}\left(\frac{-(n+1)}{4}+2.35\right) \qquad (n \geqq 10). \tag{42}$$

Combining (42) and (14) then gives

$$\operatorname{per} Y_{n+1} - \operatorname{per} Y_n < \frac{e}{8(n+1)^3}[-n+(2.35)(4)+(4)(10^{-4})-1] \tag{43}$$

for all $n \geqq 10$. Since $(2.35)(4)+(4)(10^{-4})-1<9$, (43) implies (3).

Acknowledgments. We thank Professor R. A. Askey for originally informing us about the integral representation in (8) and Professors W. Barrett and M. Lundquist for generously sharing with us the details of their derivation of (8), the implications of Pierce's conjecture that follow (2), and some useful background information on this problem. We also gratefully acknowledge some constructive comments from the referees.

REFERENCES

[1] W. Barrett and M. Lundquist, Private communication.

[2] C. L. Frenzen and R. Wong, *Asymptotic expansions for the Lebesgue constants associated with Jacobi series*, Pacific J. Math., 122 (1986), pp. 391–415.

[3] R. Grone and S. Pierce, *Permanental inequalities for correlation matrices*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 194–201.

[4] S. Pierce, *Permanents of correlation matrices*, in Current Trends in Matrix Theory, R. Grone and F. Uhlig, eds., Elsevier, North-Holland, Amsterdam, 1987.

[5] C. K. Qu and R. Wong, *Szegö's conjecture on Lebesgue constants for Legendre series*, Pacific J. Math., 135 (1988), pp. 157–188.

[6] W. Rudin, *Principles of Mathematical Analysis*, 3rd Ed., McGraw-Hill, New York, 1976.

[7] R. Wong and T. Lang, *On the points of inflection of Bessel functions of positive order*, II, Canad. J. Math., 43 (1991), pp. 628–651.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 82–91, January 1993

INTERVAL MATRICES: SINGULARITY AND REAL EIGENVALUES*

JIRI ROHN†

Abstract. This paper proves that a singular interval matrix contains a singular matrix of a very special form. This result is applied to study the real part L of the spectrum of an interval matrix. Under the assumption of sign stability of eigenvectors this paper gives a complete description of L by means of spectra of a finite subset of matrices and formulates a stability criterion for interval matrices with real eigenvalues that requires checking only two matrices for stability.

Key words. interval matrix, singular matrix, eigenvalue, stability

AMS(MOS) subject classifications. 15A03, 15A18, 65G05

1. Introduction. Let A_c and Δ be real $n \times n$ matrices with Δ nonnegative. The set of matrices

$$[A_c - \Delta, A_c + \Delta] := \{A; A_c - \Delta \leqq A \leqq A_c + \Delta\}$$

is called an interval matrix and is said to be singular if it contains a singular matrix. The problem of singularity of interval matrices is studied in [6], where in Theorem 5.1 a number of necessary and sufficient singularity conditions are given.

The purpose of the present paper is two-fold. First, we prove in Theorem 2.2 that a singular interval matrix $[A_c - \Delta, A_c + \Delta]$ contains a singular matrix A of a very special form

$$A = A_c - dT_y \Delta T_z, \tag{0}$$

where $d \in [0, 1]$ and T_z, T_y are diagonal matrices whose vectors of diagonal entries are the sign vectors of some singular vectors x and p of A and A^t, respectively. Second, in § 3 we use this theorem to study the properties of the set L of *real* eigenvalues of all matrices contained in a given interval matrix. We prove that each $\lambda \in L$ is an eigenvalue of some matrix of the form (0) (Theorem 3.2). Moreover, if $\lambda \in \partial L$, then $d = 1$ (Theorem 3.4); hence each boundary point of L is achieved at some vertex of $[A_c - \Delta, A_c + \Delta]$ (considered a polyhedron in R^{n^2}). To obtain more specific results, we introduce three assumptions imposing sign stability restrictions on eigenvectors under which we give in Theorem 3.7 a complete description of the set L as a union of at most n compact intervals whose endpoints are eigenvalues of some explicitly expressed matrices. All the results are formulated for the real part of the spectrum only, since the complex case seemingly cannot be handled by the methods used.

Stability of interval matrices has been recently extensively studied in robust control theory; see the state-of-the-art papers by Mansour [4] and Barmish [1] for detailed information. The description of the set L in Theorem 3.7 here implies a simple stability criterion in the case when only real eigenvalues are present: under the three assumptions made, an interval matrix is stable if and only if two explicitly given matrices are stable (Theorem 3.8).

We shall use the following notation. The absolute-value vector of a vector $x = (x_i)$ is defined by $|x| = (|x_i|)$. We introduce the set

$$Y = \{y \in R^n; y_j = \pm 1 \text{ for } j = 1, \ldots, n\},$$

* Received by the editors May 20, 1991; accepted for publication (in revised form) August 21, 1991.

† Faculty of Mathematics and Physics, Charles University, Malostranske nam. 25, 11800 Prague, Czechoslovakia (rohn@cspguk11.bitnet).

and for each $y \in Y$ we denote by T_y the $n \times n$ diagonal matrix with diagonal vector y. Inequalities, such as $\Delta \geqq 0$ or $\Delta > 0$, are to be understood componentwise. A^t denotes the transpose of A.

2. Singular interval matrices. Theorem 2.2 below establishes the main result of this paper; all subsequent theorems are consequences of it. It will be preceded by an auxiliary characterization of singular interval matrices.

LEMMA 2.1. *A square interval matrix* $[A_c - \Delta, A_c + \Delta]$ *is singular if and only if it satisfies*

$$|A_c x| \leqq \Delta |x|$$

for some nonzero vector x.

Proof. The assertion is an immediate consequence of the theorem by Oettli and Prager [5] that characterizes solutions of systems of linear interval equations (in our case, with zero right-hand sides); we use it here in the form given, e.g., in [6, Thm. 2.1]. □

Now we have the following main result.

THEOREM 2.2. *Let* $[A_c - \Delta, A_c + \Delta]$ *be a singular interval matrix. Then there exist* $x \neq 0$, $p \neq 0$, *and* $y, z \in Y$ *such that*

$$(A_c - dT_y\Delta T_z)x = 0, \tag{1}$$

$$(A_c - dT_y\Delta T_z)^t p = 0, \tag{2}$$

$$T_z x \geqq 0, \tag{3}$$

and

$$T_y p \geqq 0 \tag{4}$$

hold, where

$$d = \min\{\varepsilon \geqq 0;\ [A_c - \varepsilon\Delta, A_c + \varepsilon\Delta] \text{ is singular}\} \tag{5}$$

and $d \in [0, 1]$.

Comment. Notice that (3) and (4) mean that $z_j x_j \geqq 0$, $p_j y_j \geqq 0$ hold for $j = 1, \ldots, n$. Hence if all entries of x and p are nonzero, then z and y are uniquely determined and their entries are simply the signs of the respective entries of x and p. Also notice that (3) and (4) imply $T_z x = |x|$ and $T_y p = |p|$.

Proof. First observe that (1)–(4) hold trivially if A_c is singular. In this case $d = 0$; hence it suffices to take some nonzero vectors x, p that satisfy $A_c x = 0$, $A_c^t p = 0$ and choose z and y as the sign vectors of x and p; then (1)–(4) hold.

Let A_c be nonsingular, so that $d > 0$. Since the interval matrix $[A_c - d\Delta, A_c + d\Delta]$ is singular, according to the assertion (C1) of [6, Thm. 5.1], there exist $y, z \in Y$ such that $\det(A_c - dT_y\Delta T_z) \cdot \det A_c \leqq 0$. Then the continuous function φ of one real variable $\varphi(\tau) = \det(A_c - \tau d T_y \Delta T_z)$ satisfies $\varphi(0)\varphi(1) \leqq 0$; hence $\varphi(\tau_0) = 0$ for some $\tau_0 \in [0, 1]$. In view of (5) it must be $\tau_0 d \geqq d$; hence $\tau_0 = 1$, so that $A_c - dT_y\Delta T_z$ is singular and (1) holds for some $x \neq 0$. To prove the assertions (2)–(4), we shall distinguish two cases: (a) $\Delta > 0$ and (b) $\Delta \geqq 0$.

(a) Let $\Delta > 0$.

(i) To prove (3), assume to the contrary that neither $T_z x \geqq 0$ nor $T_z x \leqq 0$ holds, so that there exist $j, k \in \{1, \ldots, n\}$ such that $z_j x_j < 0$ and $z_k x_k > 0$. Take an arbitrary

$i \in \{1, \dots, n\}$. Then one of the numbers $y_i z_j x_j$, $y_i z_k x_k$ is positive and the other is negative, which gives

$$|(A_c x)_i| = d\left|\sum_h \Delta_{ih} y_i z_h x_h\right| < d(\Delta|x|)_i;$$

since i was arbitrary, it follows that $|A_c x| < d\Delta|x|$. Then we can choose a positive ε such that $\varepsilon < d$ and $|A_c x| < \varepsilon\Delta|x|$; hence the interval matrix $[A_c - \varepsilon\Delta, A_c + \varepsilon\Delta]$ is singular according to Lemma 2.1, which contradicts (5). This contradiction shows that either $T_z x \geqq 0$ or $T_z x \leqq 0$ holds. In the former case we are done, and in the latter it is sufficient to set $x := -x$ to obtain (1) and (3).

(ii) To prove (2) and (4), we first notice that since $A_c - dT_y \Delta T_z$ is singular, there exists a nonzero vector p such that

$$(A_c - dT_y\Delta T_z)^t p = (A_c^t - dT_z \Delta^t T_y)p = 0.$$

Assume that neither $T_y p \geqq 0$ nor $T_y p \leqq 0$ holds. Arguing as in (i), we obtain that $|A_c^t p| < d\Delta^t|p|$, which again gives that the interval matrix $[A_c^t - \varepsilon\Delta^t, A_c^t + \varepsilon\Delta^t]$ is singular for some $\varepsilon < d$; hence so is $[A_c - \varepsilon\Delta, A_c + \varepsilon\Delta]$, which is a contradiction. Hence either $T_y p \geqq 0$ or $T_y p \leqq 0$, so that by setting $p := -p$ if necessary we get that (2) and (4) hold also. This concludes the proof for the case of $\Delta > 0$.

(b) Let Δ be a nonnegative matrix. Let H denote the matrix of all ones and for $k = 1, 2, \dots$ define $\Delta_k = \Delta + (1/k)H$; then $\Delta_k > 0$ and each $[A_c - \Delta_k, A_c + \Delta_k]$ is singular. Hence from what has been proved under (a), it follows that for each k there exist vectors x_k, p_k (which can be normalized so that $\|x_k\|_2 = \|p_k\|_2 = 1$), and $z_k, y_k \in Y$ such that

$$(1') \qquad (A_c - d_k T_{y_k} \Delta_k T_{z_k}) x_k = 0,$$

$$(2') \qquad (A_c - d_k T_{y_k} \Delta_k T_{z_k})^t p_k = 0,$$

$$(3') \qquad T_{z_k} x_k \geqq 0,$$

$$(4') \qquad T_{y_k} p_k \geqq 0,$$

where

$$(5') \qquad d_k = \min\{\varepsilon \geqq 0;\ [A_c - \varepsilon\Delta_k, A_c + \varepsilon\Delta_k] \text{ is singular}\}.$$

First we show that $d_k \leqq d_{k+1} \leqq d$ for $k = 1, 2, \dots$. In fact, from $\Delta_k \geqq \Delta_{k+1} \geqq \Delta$ it follows that $[A_c - d_{k+1}\Delta_{k+1}, A_c + d_{k+1}\Delta_{k+1}] \subset [A_c - d_{k+1}\Delta_k, A_c + d_{k+1}\Delta_k]$, and $[A_c - d\Delta, A_c + d\Delta] \subset [A_c - d\Delta_{k+1}, A_c + d\Delta_{k+1}]$, which implies that both the interval matrices $[A_c - d_{k+1}\Delta_k, A_c + d_{k+1}\Delta_k]$ and $[A_c - d\Delta_{k+1}, A_c + d\Delta_{k+1}]$ are singular; hence $d_k \leqq d_{k+1}$ and $d_{k+1} \leqq d$ in view of (5′). Next, since Y is finite, there exists a constant subsequence of the sequence $\{(z_k, y_k)\}_{k=1}^{\infty}$, i.e., $z_k = z$, $y_k = y$ for infinitely many k. Let us choose another subsequence of this subsequence along which x_k, p_k converge to some x, p (this is possible since $\{x_k\}$, $\{p_k\}$ are confined to the compact unit sphere; hence $x \neq 0$ and $p \neq 0$). Then taking limits in (1′)–(4′) we obtain

$$(A_c - \bar{d} T_y \Delta T_z) x = 0,$$

$$(A_c - \bar{d} T_y \Delta T_z)^t p = 0,$$

$$T_z x \geqq 0,$$

$$T_y p \geqq 0,$$

where $\bar{d} = \lim_{k\to\infty} d_k \leqq d$. Since the matrix $A_c - \bar{d}T_y\Delta T_z$ is singular and belongs to the interval matrix $[A_c - \bar{d}\Delta, A_c + \bar{d}\Delta]$, we have in light of (5) that $d \leqq \bar{d}$; hence $\bar{d} = d \in [0, 1]$, so that (1)–(4) hold and the proof is complete. □

Next, we formulate some direct consequences of Theorem 2.2. First, we show that there exists a singular matrix in a "normal form."

COROLLARY 2.3. *Let* $[A_c - \Delta, A_c + \Delta]$ *be singular. Then it contains a singular matrix of the form*

$$A_c - dT_y\Delta T_z, \tag{6}$$

where $y, z \in Y$ *and* $d \in [0, 1]$.

Proof. This is an obvious consequence of the assertions (1) and (5) of Theorem 2.2. □

The result can also be given the following geometric formulation.

COROLLARY 2.4. *Let* $[A_c - \Delta, A_c + \Delta]$ *be singular. Then it contains a singular matrix belonging to a segment connecting* A_c *with some vertex of* $[A_c - \Delta, A_c + \Delta]$ (*considered a polyhedron in* R^{n^2}).

Proof. For the singular matrix A from (6) we have

$$A = (1 - d)A_c + d(A_c - T_y\Delta T_z),$$

where $d \in [0, 1]$; hence A belongs to the segment connecting A_c with the matrix $A_c - T_y\Delta T_z$, which is a vertex of $[A_c - \Delta, A_c + \Delta]$ since $(A_c - T_y\Delta T_z)_{ij} = (A_c - \Delta)_{ij}$ if $y_iz_j = 1$ and $(A_c - T_y\Delta T_z)_{ij} = (A_c + \Delta)_{ij}$ if $y_iz_j = -1$. □

COROLLARY 2.5. *Let* $[A_c - \Delta, A_c + \Delta]$ *be singular. Then there exists an* $x \neq 0$ *such that*

$$|A_cx| = d\Delta|x|$$

holds, where d *is given by* (5).

Comment. The assertion is stronger than that of Lemma 2.1; it shows that the inequality holds "uniformly."

Proof. From (1) and (3) we have $|A_cx| = |dT_y\Delta T_zx| = d|\Delta|x|| = d\Delta|x|$. □

3. Real eigenvalues of an interval matrix. In this section we shall apply Theorem 2.2 to study the set of real eigenvalues of an interval matrix $[A_c - \Delta, A_c + \Delta]$ given by

$$L = \{\lambda \in R^1; Ax = \lambda x \text{ for some } A \in [A_c - \Delta, A_c + \Delta], x \neq 0\}.$$

Obviously, L is compact since each $\lambda \in L$ can be written as $\lambda = x^tAx$ for some $A \in [A_c - \Delta, A_c + \Delta]$ and some x with $\|x\|_2 = 1$. In the sequel we shall use the following result.

LEMMA 3.1. $\lambda \in L$ *if and only if the interval matrix*

$$[(A_c - \lambda I) - \Delta, (A_c - \lambda I) + \Delta] \tag{7}$$

is singular.

Proof. If $\lambda \in L$, then $(A - \lambda I)x = 0$ for some $x \neq 0$, where $A - \lambda I$ belongs to the interval matrix (7), which is then singular. Conversely, if (7) is singular, then it contains a matrix A with $Ax = 0$, $x \neq 0$. Then $A + \lambda I \in [A_c - \Delta, A_c + \Delta]$ and $(A + \lambda I)x = \lambda x$; hence $\lambda \in L$. □

We shall first show that each $\lambda \in L$ is an eigenvalue of a matrix of a special form.

THEOREM 3.2. *Let $\lambda \in L$. Then there exist $x \neq 0$, $p \neq 0$, $y, z \in Y$, and $d \in [0, 1]$ such that*

$$(A_c - dT_y \Delta T_z)x = \lambda x, \tag{8a}$$

$$(A_c - dT_y \Delta T_z)^t p = \lambda p, \tag{8b}$$

$$T_z x \geqq 0, \tag{8c}$$

$$T_y p \geqq 0. \tag{8d}$$

Proof. The assertion is a direct consequence of Theorem 2.2 applied according to Lemma 3.1 to the interval matrix (7). □

Let us introduce, as in [6], the matrices

$$A_{yz} = A_c - T_y \Delta T_z$$

for $y, z \in Y$. Obviously, $A_{yz} \in [A_c - \Delta, A_c + \Delta]$ for each $y, z \in Y$.

COROLLARY 3.3. *Let $\lambda \in L$. Then λ is an eigenvalue of a matrix belonging to a segment connecting A_c with some matrix A_{yz} for $y, z \in Y$.*

Proof. The proof follows from Theorem 3.2. □

Hence all the real eigenvalues of $[A_c - \Delta, A_c + \Delta]$ are achieved at matrices belonging to a finite number of segments, i.e., to a set of measure zero if $n > 1$; see Hollot and Bartlett [3] for a similar result using the edges of $[A_c - \Delta, A_c + \Delta]$.

Now we shall show that the boundary points of L are eigenvalues of the matrices A_{yz}.

THEOREM 3.4. *Let $\lambda \in \partial L$. Then there exist $x \neq 0$, $p \neq 0$, and $y, z \in Y$ such that*

$$A_{yz} x = \lambda x, \tag{9a}$$

$$A_{yz}^t p = \lambda p, \tag{9b}$$

$$T_z x \geqq 0, \tag{9c}$$

$$T_y p \geqq 0. \tag{9d}$$

Proof. As in the proof of Theorem 2.2, we shall consider separately two cases: (a) $\Delta > 0$ and (b) $\Delta \geqq 0$.

(a) Let $\Delta > 0$. Since L is compact, we have $\partial L \subset L$; hence (8a)–(8d) hold for some $x \neq 0$, $p \neq 0$, $y, z \in Y$ and $d \in [0, 1]$. We shall prove that $d = 1$. Assume to the contrary that $d < 1$. Then from Theorem 3.2 we have

$$|(A_c - \lambda I)x| = d\Delta |x| < \Delta |x|.$$

Hence there exists an $\varepsilon > 0$ such that each $\lambda' \in (\lambda - \varepsilon, \lambda + \varepsilon)$ satisfies

$$|(A_c - \lambda' I)x| < \Delta |x|.$$

Therefore, $[(A_c - \lambda' I) - \Delta, (A_c - \lambda' I) + \Delta]$ is singular according to Lemma 2.1. This implies $\lambda' \in L$ by Lemma 3.1. Hence λ is an interior point of L, which contradicts the assumption that $\lambda \in \partial L$. Thus $d = 1$, so that (8a)–(8d) take on the form of (9a)–(9d).

(b) Let $\Delta \geqq 0$. As in the second part of the proof of Theorem 2.2, for $k = 1, 2, \ldots$ define $\Delta_k = \Delta + (1/k)H$, where H is the matrix of all ones, and let L_k be the set of real eigenvalues of the interval matrix $[A_c - \Delta_k, A_c + \Delta_k]$, so that $L \subset L_k$. For each $k = 1, 2, \ldots$ take a $\lambda_k \in \partial L_k$ that satisfies

$$|\lambda_k - \lambda| = \min \{ |\tilde{\lambda} - \lambda|; \tilde{\lambda} \in \partial L_k \}.$$

We shall prove that $\lambda_k \to \lambda$. Assume this is not the case; then there exists an $\varepsilon > 0$ and a subsequence $\{k_j\}$ such that $|\lambda_{k_j} - \lambda| \geqq \varepsilon$ for $j = 1, 2, \ldots$, which in view of the definition of λ_{k_j} gives that $(\lambda - \varepsilon, \lambda + \varepsilon) \subset L_{k_j}$ for each j. Hence for each $\lambda' \in (\lambda - \varepsilon, \lambda + \varepsilon)$ and each $j = 1, 2, \ldots$ there exists a matrix $A_{k_j} \in [A_c - \Delta_{k_j}, A_c + \Delta_{k_j}] \subset [A_c - \Delta_1, A_c + \Delta_1]$ and a vector x_{k_j} with $\|x_{k_j}\|_2 = 1$ such that $A_{k_j} x_{k_j} = \lambda' x_{k_j}$. Taking the limit, we obtain $A_0 x_0 = \lambda' x_0$ for some $A_0 \in [A_c - \Delta, A_c + \Delta]$ and $x_0 \neq 0$. Hence $\lambda' \in L$. This gives $(\lambda - \varepsilon, \lambda + \varepsilon) \subset L$, contrary to $\lambda \in \partial L$. Hence $\lambda_k \to \lambda$. Now, since $\lambda_k \in \partial L_k$ and $\Delta_k > 0$, by applying the result proved in part (a) we obtain that for each $k = 1, 2, \ldots$ there exist $y_k, z_k \in Y$ and vectors x_k, p_k with $\|x_k\|_2 = \|p_k\|_2 = 1$ such that

$$(A_c - T_{y_k}\Delta_k T_{z_k})x_k = \lambda_k x_k,$$

$$(A_c - T_{y_k}\Delta_k T_{z_k})^t p_k = \lambda_k p_k,$$

$$T_{z_k} x_k \geqq 0,$$

$$T_{y_k} p_k \geqq 0$$

hold. Choosing a subsequence along which y_k, z_k remain constant and x_k, p_k converge, we obtain (9a)–(9d), which completes the proof. □

Theorems 3.2 and 3.4 were quite general; to achieve more specific results about the structure of L, we now introduce some assumptions.

Assumption 1. Each $A \in [A_c - \Delta, A_c + \Delta]$ has exactly m real eigenvalues ($1 \leqq m \leqq n$) numbered in such a way that $\lambda_1(A) < \cdots < \lambda_m(A)$.

Then we can define the sets

$$L_i = \{\lambda_i(A); A \in [A_c - \Delta, A_c + \Delta]\}$$

for $i = 1, \ldots, m$.

Assumption 2. $\bar{L}_i \cap \bar{L}_j = \varnothing$ for each $i \neq j$, $i, j \in \{1, \ldots, m\}$ (where the bar denotes closure).

Next we shall assume a sign pattern constancy of the eigenvectors.

Assumption 3. For each $i \in \{1, \ldots, m\}$ there exist vectors $z_i, y_i \in Y$ such that each right eigenvector x (left eigenvector p) pertaining to the ith real eigenvalue of some $A \in [A_c - \Delta, A_c + \Delta]$ satisfies either $T_{z_i} x > 0$ or $T_{z_i} x < 0$ ($T_{y_i} p > 0$ or $T_{y_i} p < 0$).

We formulate the third assumption in this way because $-x$ and $-p$ are also a right eigenvector and left eigenvector, respectively. Notice that $-z_i$ and $-y_i$ also possess the required property.

COROLLARY 3.5. *Let Assumptions* 1–3 *be satisfied, and let* $i \in \{1, \ldots, m\}$. *Then each* $\lambda \in L_i$ *is the ith real eigenvalue of some matrix belonging to the segment connecting* $A_{y_i z_i}$ *with* A_{-y_i, z_i}.

Proof. According to Theorem 3.2 there exist x, p, y, z, and d satisfying (8a)–(8d). Hence λ is an eigenvalue of the matrix $A = A_c - dT_y \Delta T_z$. If $\lambda = \lambda_j(A)$ for some $j \neq i$, then $\lambda \in L_i \cap L_j$, which contradicts Assumption 2. Hence $\lambda = \lambda_i(A)$, which, according to Assumption 3 in conjunction with (8c) and (8d), means that $z = \pm z_i$ and $y = \pm y_i$. Hence either

$$A = A_c - dT_{y_i}\Delta T_{z_i} = \tfrac{1}{2}(1 + d)A_{y_i z_i} + \tfrac{1}{2}(1 - d)A_{-y_i, z_i}$$

or

$$A = A_c - dT_{-y_i}\Delta T_{z_i} = \tfrac{1}{2}(1 - d)A_{y_i z_i} + \tfrac{1}{2}(1 + d)A_{-y_i, z_i}$$

with $d \in [0, 1]$. In both cases A belongs to the segment connecting $A_{y_i z_i}$ with A_{-y_i, z_i}. □

An interval matrix $[A_c - \Delta, A_c + \Delta]$ is called symmetric if both A_c and Δ are symmetric. In this case, if $A \in [A_c - \Delta, A_c + \Delta]$, then $A^t \in [A_c - \Delta, A_c + \Delta]$ also, but generally $[A_c - \Delta, A_c + \Delta]$ can contain nonsymmetric matrices. However, we have the following result.

COROLLARY 3.6. *Let a symmetric interval matrix* $[A_c - \Delta, A_c + \Delta]$ *satisfy Assumptions* 1–3. *Then each* $\lambda \in L$ *is an eigenvalue of some symmetric matrix in* $[A_c - \Delta, A_c + \Delta]$.

Proof. Since each left eigenvector of A_c is also a right eigenvector in this case, it follows from Assumption 3 that $y_i = \pm z_i$ for $i = 1, \ldots, m$. If $\lambda \in L$, then $\lambda \in L_i$ for some i and Corollary 3.5 implies that λ is an eigenvalue of a matrix of the form $A_c - dT_{z_i}\Delta T_{z_i}$ for $d \in [-1, 1]$, which is obviously symmetric. □

Now we are ready to describe the structure of L. The following theorem is a generalization of [7, Thm. 3], where a similar result is proved for interval matrices with Δ of rank one, i.e., of the form $\Delta = qp^t$ for some positive vectors q and p, whereas now Δ can be an arbitrary nonnegative matrix; see also Deif [2].

THEOREM 3.7. *Let an interval matrix* $[A_c - \Delta, A_c + \Delta]$ *satisfy Assumptions* 1–3. *Then*

$$L = \bigcup_{i=1}^{m} L_i,$$

where for each $i \in \{1, \ldots, m\}$ *we have*

$$L_i = [\underline{\lambda}_i, \bar{\lambda}_i] \tag{10}$$

with

$$\underline{\lambda}_i = \min\{\lambda_i(A_{y_iz_i}), \lambda_i(A_{-y_i,z_i})\} \tag{11}$$

and

$$\bar{\lambda}_i = \max\{\lambda_i(A_{y_iz_i}), \lambda_i(A_{-y_i,z_i})\}. \tag{12}$$

Comment. L is thus completely determined by the real components of the spectra of $2m$ matrices $A_{y_iz_i}$, A_{-y_i,z_i} $(i = 1, \ldots, m)$.

Proof. The assertion for L is simply a consequence of the definition of the L_i's; therefore, we need to prove only (10)–(12). These are obvious if $\Delta = 0$. Assume $\Delta \neq 0$, and let $i \in \{1, \ldots, m\}$, $\lambda \in \partial L_i$. Then $\lambda \in \partial L$, or else Assumption 2 would be violated. Hence, according to Theorem 3.4, λ is an eigenvalue of some A_{yz}, and in view of Assumption 2, $\lambda = \lambda_i(A_{yz})$. From (9c) and (9d) we can infer, as in the proof of Corollary 3.5, that either $\lambda = \underset{\sim}{\lambda}_i$ or $\lambda = \tilde{\lambda}_i$, where we have denoted $\underset{\sim}{\lambda}_i = \lambda_i(A_{-y_i,z_i})$ and $\tilde{\lambda}_i = \lambda_i(A_{y_iz_i})$. Hence L_i has at most two boundary points. Furthermore, let x and p be a right eigenvector and a left eigenvector to $\underset{\sim}{\lambda}_i$ and $\tilde{\lambda}_i$, respectively, such that $T_{z_i}x = |x| > 0$ and $T_{y_i}p = |p| > 0$. Then we have

$$(A_c + T_{y_i}\Delta T_{z_i})x = \underset{\sim}{\lambda}_i x$$

and

$$(A_c - T_{y_i}\Delta T_{z_i})^t p = \tilde{\lambda}_i p.$$

Premultiplying the first equation by p^t and the second by x^t and subtracting, we obtain

$$2|p|^t\Delta|x| = (\underset{\sim}{\lambda}_i - \tilde{\lambda}_i)p^t x.$$

Since $|p| > 0$, $|x| > 0$, and $\Delta \neq 0$, this shows that $\underset{\sim}{\lambda}_i \neq \tilde{\lambda}_i$. In a similar manner we can obtain $\underset{\sim}{\lambda}_i \neq \lambda_i(A_c)$ and $\tilde{\lambda}_i \neq \lambda_i(A_c)$. To sum up, we have proved that the compact set L_i consists of at least three different points and has at most two boundary points $\underset{\sim}{\lambda}_i$ and $\tilde{\lambda}_i$; hence L_i is a compact interval whose endpoints are the two boundary points, i.e.,

$$L_i = [\underline{\lambda}_i, \bar{\lambda}_i],$$

where

$$\underline{\lambda}_i = \min \{\underset{\sim}{\lambda}_i, \tilde{\lambda}_i\}$$

and

$$\bar{\lambda}_i = \max \{\underset{\sim}{\lambda}_i, \tilde{\lambda}_i\}. \qquad \square$$

Notice that if the interval matrix is symmetric, then the extremal eigenvalues $\underline{\lambda}_i$, $\bar{\lambda}_i$ are achieved at symmetric matrices $A_{z_i z_i}$ and A_{-z_i, z_i} $(i = 1, \dots, m)$.

The above result has an implication for stability of interval matrices with real eigenvalues. A square matrix is called stable [4] if all its eigenvalues are placed in the open left half of the complex plane. An interval matrix $[A_c - \Delta, A_c + \Delta]$ is called stable if each $A \in [A_c - \Delta, A_c + \Delta]$ is stable. We have the following characterization.

THEOREM 3.8. *Let an interval matrix* $[A_c - \Delta, A_c + \Delta]$ *satisfy Assumptions* 1–3 *with* $m = n$. *Then it is stable if and only if* $A_{y_n z_n}$ *and* A_{-y_n, z_n} *are stable.*

Proof. Since $m = n$, all the eigenvalues are real. Therefore, for each $A \in [A_c - \Delta, A_c + \Delta]$ and each $i \in \{1, \dots, n\}$ we have

$$\lambda_i(A) < \lambda_n(A) \leqq \bar{\lambda}_n = \max \{\lambda_n(A_{y_n z_n}), \lambda_n(A_{-y_n, z_n})\} < 0.$$

Hence A is stable. The "only if" part is obvious. $\square$

4. Real eigenvectors of an interval matrix. In this section we give a characterization of real eigenvectors of matrices belonging to a given interval matrix. In contrast to the eigenvalue case, the situation is much simpler here because it turns out that real eigenvectors can be characterized by a verifiable necessary and sufficient condition.

THEOREM 4.1. *A nonzero real vector* x *is an eigenvector of some matrix in* $[A_c - \Delta, A_c + \Delta]$ *if and only if the matrix*

$$X = |x| \cdot |x|^t$$

satisfies

$$(13) \qquad (T_z A_c T_z - \Delta) X \leqq X (T_z A_c T_z + \Delta)^t,$$

where the vector z *is given by* $z_j = 1$ *if* $x_j \geqq 0$ *and* $z_j = -1$ *otherwise* $(j = 1, \dots, n)$.

Proof. For the sake of brevity, let us denote $A_z = T_z A_c T_z$. Then (13) becomes

$$(14) \qquad (A_z - \Delta) X \leqq X (A_z + \Delta)^t.$$

"Only if": Let x be a real eigenvector of a matrix $A \in [A_c - \Delta, A_c + \Delta]$, so that $Ax = \lambda x$ for some real λ. Since $T_z x = |x|$ by the definition of z, we have

$$|(A_z - \lambda I)|x|| = |A_c x - \lambda x| = |(A_c - A)x| \leqq \Delta |x|,$$

which implies that

$$(15) \qquad (A_z - \Delta)|x| \leqq \lambda |x| \leqq (A_z + \Delta)|x|.$$

Premultiplying the left-hand inequality by $|x|^t$, we obtain

$$(A_z - \Delta)X \leqq \lambda X,$$

and then, transposing the right-hand inequality in (15) and premultiplying the result by $|x|$, we have

$$\lambda X \leqq X(A_z + \Delta)^t,$$

which together give (14).

"If": Conversely, let x be a nonzero real vector such that the matrix $X = |x| \cdot |x|^t$ satisfies (14). Then for each $i, j \in \{1, \ldots, n\}$ we have

$$((A_z - \Delta)|x|)_i |x_j| \leqq |x_i|((A_z + \Delta)|x|)_j. \tag{16}$$

Hence for each i, j with $x_i \neq 0$, $x_j \neq 0$ it holds that

$$\frac{((A_z - \Delta)|x|)_i}{|x_i|} \leqq \frac{((A_z + \Delta)|x|)_j}{|x_j|}$$

and, consequently,

$$\max_{x_i \neq 0} \frac{((A_z - \Delta)|x|)_i}{|x_i|} \leqq \min_{x_j \neq 0} \frac{((A_z + \Delta)|x|)_j}{|x_j|}.$$

Hence there exists a λ satisfying

$$\max_{x_i \neq 0} \frac{((A_z - \Delta)|x|)_i}{|x_i|} \leqq \lambda \leqq \min_{x_j \neq 0} \frac{((A_z + \Delta)|x|)_j}{|x_j|}. \tag{17}$$

We shall show that (λ, x) is an eigenpair of some matrix in $[A_c - \Delta, A_c + \Delta]$. Let $k \in \{1, \ldots, n\}$. If $x_k \neq 0$, then (17) gives

$$\frac{((A_z - \Delta)|x|)_k}{|x_k|} \leqq \lambda \leqq \frac{((A_z + \Delta)|x|)_k}{|x_k|}.$$

Hence

$$((A_z - \Delta)|x|)_k \leqq \lambda |x_k| \tag{18}$$

and

$$\lambda |x_k| \leqq ((A_z + \Delta)|x|)_k \tag{19}$$

hold. If $x_k = 0$, select an m with $x_m \neq 0$. Then from the inequality (16) applied to $i = k, j = m$ we obtain (18), and similarly applying it to $i = m, j = k$, we get (19). Hence (18) and (19) are valid for each k, which gives that

$$|(A_c - \lambda I)x| = |(A_z - \lambda I)|x|| \leqq \Delta |x|.$$

Then Lemmas 2.1 and 3.1 imply the existence of a matrix $A \in [A_c - \Delta, A_c + \Delta]$ such that $(A - \lambda I)x = 0$. Hence $\lambda \in L$, and x is an eigenvector of A. □

Acknowledgments. The author wishes to thank the referees and the Reviewing Editor G. Strang for comments that helped to improve the text of this paper.

REFERENCES

[1] B. R. BARMISH, *New tools for robustness analysis*, in Proc. 27th Conference on Decision and Control, Austin, TX, 1988, pp. 1–6.

[2] A. DEIF, *The interval eigenvalue problem*, Z. Angew. Math. Mech., 71 (1991), pp. 61–64.

[3] C. V. HOLLOT AND A. C. BARTLETT, *On the eigenvalues of interval matrices*, in Proc. 26th Conference on Decision and Control, Los Angeles, CA, 1987, pp. 794–799.

[4] M. MANSOUR, *Robust stability of interval matrices*, in Proc. 28th Conference on Decision and Control, Tampa, FL, 1989.

[5] W. OETTLI AND W. PRAGER, *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math., 6 (1964), pp. 405–409.

[6] J. ROHN, *Systems of linear interval equations*, Linear Algebra Appl., 126 (1989), pp. 39–78.

[7] ———, *Real eigenvalues of an interval matrix with rank one radius*, Z. Angew. Math. Mech., 70 (1990), pp. T562–T563.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 92-112, January 1993

008

MATRIX GENERALIZATIONS OF A MOMENT PROBLEM THEOREM I. THE HERMITIAN CASE*

MIRON TISMENETSKY†

Abstract. The basic results of Caratheodory and Achiezer and Krein regarding certain representations of elements of a positive definite Toeplitz matrix are generalized to the case of block Toeplitz matrices. It is shown that these results are closely related to some factorizations of block Toeplitz matrices, matrix polynomials, and indefinite scalar products.

Key words. Toeplitz matrix, factorization, indefinite scalar product, matrix polynomial, matrix equation

AMS(MOS) subject classifications. 15A23, 15A57, 47A68

0. Introduction.

The main objective of this work is a generalization of the following Caratheodory–Achiezer–Krein theorem.

THEOREM 0.1 (see [1]–[3]). *If $A = [a_{i-j}]_{i,j=1}^{l}$ $(a_{-k} = \bar{a}_k)$ is a Hermitian positive definite Toeplitz matrix, then to every complex number z such that*

$$\det \begin{bmatrix} a_0 & \bar{a}_1 & \cdots & \bar{a}_{l-1} & \bar{z} \\ a_1 & a_0 & \cdots & \cdot & \bar{a}_{l-1} \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ a_{l-1} & \cdot & \cdots & \cdot & \bar{a}_1 \\ z & a_{l-1} & \cdots & a_1 & a_0 \end{bmatrix} = 0, \tag{0.1}$$

there corresponds one and only one canonical representation

$$a_k = \sum_{j=1}^{l} \rho_j \mu_j^k \qquad (k = 0, \pm 1, \cdots \pm (l-1)), \tag{0.2}$$

in which $\rho_j > 0$, $|\mu_j| = 1$ $(j = 1, 2, \ldots, l)$ and all μ_j's are distinct and such that

$$z = \sum_{j=1}^{l} \rho_j \mu_j^l. \tag{0.3}$$

Conversely, to every representation (0.2) *of the indicated type there corresponds by formula* (0.3) *some z satisfying* (0.1).

Theorem 0.1 has its roots in Caratheodory's investigations [2], [3]. The results stirred up much interest, as can be seen from the many proofs produced by eminent mathematicians of that time (cf. Schur [15], Frobenius [7], and Szego [16]). As stated, Theorem 0.1 appears in the Achiezer–Krein paper [1], which provides a comprehensive treatment of this and related problems, and their connections to the truncated trigonometric moment problem, polynomials orthogonal on the unit circle, and other topics.

* Received by the editors June 18, 1990; accepted for publication (in revised form) May 8, 1991.

† IBM Haifa Scientific Center, Haifa, Israel (timor@haifasc3.vnet.ibm.com) and Institute of Supercomputing and Applied Mathematics, IBM Scientific Center, D-6900 Heidelberg, Germany.

It is easily observed that the representation (0.2) is equivalent to the following factorization of matrix A:

$$
A = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \mu_1 & \mu_2 & \cdots & \mu_l \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mu_1^{l-1} & \mu_2^{l-1} & \cdots & \mu_l^{l-1} \end{bmatrix} \begin{bmatrix} \rho_1 & 0 & \cdots & \cdot & 0 \\ 0 & \rho_2 & \cdots & \cdot & \cdot \\ \cdot & 0 & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & 0 & \cdot \\ \cdot & \cdot & \cdots & \rho_{l-1} & 0 \\ 0 & \cdot & \cdots & 0 & \rho_l \end{bmatrix}
\tag{0.4}
$$

$$
\times \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \mu_1 & \mu_2 & \cdots & \mu_l \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mu_1^{l-1} & \mu_2^{l-1} & \cdots & \mu_l^{l-1} \end{bmatrix}^*,
$$

in which the condition $\mu_j \neq \mu_i$ for $j \neq i$ means that the Vandermonde matrix on the left is nonsingular. The unimodularity of μ_j's (that is, $|\mu_j| = 1$) forces the product of the matrices on the right in (0.4) to be Toeplitz. The condition (0.3) shows that, along with the factorization (0.4), there exists a factorization of similar type of the extended $(l+1) \times (l+1)$ Toeplitz matrix occurring in (0.1).

In this first part of the work our generalizations are concerned with the case when A is a block Toeplitz matrix with $n \times n$ blocks and is nonsingular Hermitian, but not necessarily positive definite. Hence the Caratheodory–Achiezer–Krein theorem is generalized also in the scalar case. The case of an arbitrary $n \times n$ block Toeplitz matrix is treated in the second part of the work.

Viewing the numbers μ_j appearing in (0.2) as roots of a monic polynomial $L(\lambda)$ of degree l, we derive generalizations of Theorem 0.1 in terms of the Gohberg–Lancaster–Rodman spectral theory of matrix polynomials developed in [9] and subsequent works. More precisely, let

$$
L(\lambda) = L_0 + \lambda L_1 + \cdots + \lambda^{l-1} L_{l-1} + \lambda^l L_l \qquad (L_k \in \mathbf{C}^{n \times n},\ k = 0, 1, \ldots, l)
$$

be an $n \times n$ *monic* matrix polynomial (that is, $L_l = I$, the $n \times n$ identity matrix). The pair of matrices (X, T), where X is $n \times ln$ and T is $ln \times ln$ is said to be a (*right*) *standard pair* of $L(\lambda)$ if the $ln \times ln$ matrix

$$
\operatorname{col}\,(XT^j)_{j=0}^{l-1} \doteq \begin{bmatrix} X \\ XT \\ \cdot \\ \cdot \\ \cdot \\ XT^{l-1} \end{bmatrix}
$$

is nonsingular and the equation

$$
L_0 X + L_1 XT + \cdots L_{l-1} XT^{l-1} + XT^l = 0
$$

holds. The matrix T is called a *linearization* of $L(\lambda)$, and can be characterized as a matrix similar to the (*block*) *companion matrix*

$$(0.5) \qquad C_L = \begin{bmatrix} 0 & I & \cdot & \cdots & 0 \\ \cdot & \cdot & I & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdot & \cdots & I \\ -L_0 & -L_1 & \cdot & \cdots & -L_{l-1} \end{bmatrix}$$

associated with the matrix polynomial $L(\lambda)$.

Clearly, the pair of matrices (X, C_L), where the matrix $X = [I\, 0 \cdots 0]$ is $n \times ln$, is an example of a (right) standard pair for $L(\lambda)$. Another standard pair, called a *Jordan pair* (see [9] for definition), provides a complete explicit description of the spectral properties of the given matrix polynomial.

Let R denote a nonsingular Hermitian matrix. A nonsingular matrix T is said to be *quasi-unitary with respect to the indefinite scalar product generated by* R or, briefly, R-*unitary* if $T^*RT = R$. It turns out that this notion is intimately connected to the Caratheodory–Achiezer–Krein-type decompositions of block Toeplitz matrices.

The following result presents a weak version of our main Theorem 3.1 concerning a generalization of the Caratheodory–Achiezer–Krein theorem for nonsingular Hermitian block Toeplitz matrices.

THEOREM 0.2. *Let A denote a nonsingular Hermitian block Toeplitz matrix*:

$$A = [A_{i-j}]_{i,j=1}^{l} \qquad (A_k \in \mathbf{C}^{n \times n}, A_{-k} = A_k^*, k = 0, \pm 1, \ldots, \pm(l-1)).$$

Then to every $n \times n$ matrix Z such that

$$(0.6) \qquad \operatorname{rank} \begin{bmatrix} A_0 & A_1^* & \cdots & A_{l-1}^* & Z^* \\ A_1 & A_0 & \cdots & \cdot & A_{l-1}^* \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ A_{l-1} & \cdot & \cdots & \cdot & A_1^* \\ Z & A_{l-1} & \cdots & A_1 & A_0 \end{bmatrix} = \operatorname{rank} A,$$

there corresponds a unique set of integers $r_1, r_2, \ldots, r_p$ ($\sum r_k = ln$) such that the $n \times n$ entries of A admit a representation

$$(0.7) \qquad A_k = \sum_{j=1}^{p} X_j T_j^{*k} R_j X_j^* \qquad (k = 0, \pm 1, \ldots, \pm(l-1))$$

having the following properties:

(a) *the matrices R_j are $r_j \times r_j$ and nonsingular Hermitian*;

(b) *the $r_j \times r_j$ matrices T_j are R_j-unitary and hence have eigenvalues symmetric relative to the unit circle*;

(c) *the matrices X_j are $n \times r_j$ and such that the matrix*

$$\tilde{X} \doteq \begin{bmatrix} X_1 & X_2 & \cdots & X_p \\ X_1 T_1 & X_2 T_2 & \cdots & X_p T_p \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ X_1 T_1^{l-1} & X_2 T_2^{l-1} & \cdots & X_p T_p^{l-1} \end{bmatrix}$$

is square and nonsingular;

(d) *there exists a representation*

$$Z = \sum_{j=1}^{p} X_j T_j^{*l} R_j X_j^* . \tag{0.8}$$

Conversely, for every representation (0.7), *the matrix Z computed by formula* (0.8) *satisfies* (0.6).

The matrix $\tilde{X}$ above can be naturally viewed as a *generalized Vandermonde matrix* and briefly written as

$$\tilde{X} = \operatorname{col}\,(XT^j)_{j=0}^{l-1},$$

where

$$X = [X_1 X_2 \cdots X_p], \qquad T = \operatorname{diag}\,[T_1, T_2, \ldots, T_p],$$

and diag $[T_1, T_2, \ldots, T_p]$ stands for the block diagonal matrix with the matrices T_1, $T_2, \ldots, T_p$ on its main (block) diagonal.

With this notation, the statement of Theorem 0.2 can be reformulated as the existence of a factorization

$$A = \tilde{X} R \tilde{X}^*,$$

where

$$R = \operatorname{diag}\,[R_1, R_2, \ldots, R_p].$$

The matrices X and T above are $n \times ln$ and $ln \times ln$, respectively, and such that det $\tilde{X} \neq 0$. Hence (cf. [9]) there exists a monic $n \times n$ matrix polynomial $L(\lambda) = \sum_{j=0}^{l-1} \lambda^j L_j + \lambda^l I$ having (X, T) as its (right) standard pair. Furthermore, it turns out that the coefficients of this polynomial form a solution of a (block) *Yule–Walker-type equation*

$$-[L_0 L_1 \cdots L_{l-1}]A = [Z A_{l-1} A_{l-2} \cdots A_1], \tag{0.9}$$

where the matrix Z generates a Hermitian *singular extension* $\hat{A}$ of A defined as an $(l+)n \times (l+1)n$ Hermitian block Toeplitz matrix appearing on the left in (0.6), and such that

$$\operatorname{rank} \hat{A} = \operatorname{rank} A.$$

To specify an extension of A, we refer to the matrix above as to the (Hermitian) *Z-extension* of A.

The existence of Hermitian singular extensions for positive definite block Toeplitz matrices has been established in the first part of [6], where it is shown, in particular, that the set $\{Z\}$ of all matrices producing singular Hermitian extensions of a Hermitian nonnegative definite matrix A constitutes a matrix sphere in $\mathbf{C}^{n \times n}$. We also note that the characterization of the positive definite case (Theorem 3.2 below) yields a description of the block structure which is qualitatively different from that in [6].

A close connection between the matrix polynomials $L(\lambda)$ derived from (0.9) and homogeneous Stein equations in companion matrices is indicated. In particular (see Theorem 2.1 for the complete statement), it is shown that, for any nonsingular Hermitian block Toeplitz matrix A, (0.9) with Z producing a singular Hermitian extension of A yields the following equation in the companion matrices:

$$C_L A C_L^* = A,$$

and vice versa. The latter shows that the matrix C_L^* is A-unitary and, in particular, that the spectrum of the companion matrix for $L(\lambda)$, as well as that of $L(\lambda)$ itself, is symmetric with respect to the unit circle (cf. [10]). Recall that the spectrum of a monic matrix polynomial $L(\lambda)$ is defined as the set $\sigma(L) = \{\mu \in \mathbf{C}\colon \det L(\mu) = 0\}$. The members of $\sigma(L)$ are referred to as *eigenvalues* of the monic matrix polynomial $L(\lambda)$.

The results obtained in this work (especially for positive definite matrices) are closely related to the truncated trigonometric matrix moment problem, the theory of polynomials orthogonal on the unit circle, the Caratheodory problem, the prediction theory of multivariate stationary sequences, etc. (see [1], [4], [6], and [8] and references quoted therein). An investigation of these relations, as well as a further development of the approach adopted in this paper, will be carried out in a subsequent work. Block Hankel analogs of the Caratheodory–Achiezer–Krein theorem are obtained in [17].

1. First observations. We start with observations concerning products of three matrices.

THEOREM 1.1. *Let X and T be any matrices of sizes $n \times r$ and $r \times r$, respectively. Then for any $r \times r$ matrix R satisfying the condition $TRT^* = R$ and any integer l, the product*

$$(1.1) \qquad A \doteq \begin{bmatrix} X \\ XT \\ \cdot \\ \cdot \\ \cdot \\ XT^{l-1} \end{bmatrix} R \begin{bmatrix} X \\ XT \\ \cdot \\ \cdot \\ \cdot \\ XT^{l-1} \end{bmatrix}^*$$

is an $ln \times ln$ block Toeplitz matrix with $n \times n$ blocks. If the matrix $\operatorname{col}(XT^j)_{j=0}^{l-2}$ *is left invertible, then, conversely, the availability of* (1.1) *yields the condition $TRT^* = R$.*

Proof. Matrix A in (1.1) is block Toeplitz if and only if

$$(1.2) \qquad XT^jRT^{*k}X^* = XT^{j-1}RT^{*k-1}X^*$$

for $j, k = 1, 2, \ldots, l-1$. Or, which is equivalent,

$$\operatorname{col}(XT^j)_{j=0}^{l-2}(R - TRT^*)[\operatorname{col}(XT^j)_{j=0}^{l-2}]^* = 0.$$

Hence the assertion of the theorem. □

A particularly important form of the product in (1.1) is presented below for the case when X, T, and R can be decomposed as follows:

$$X = [X_1X_2\cdots X_p], \quad T = \operatorname{diag}[T_1, T_2, \ldots, T_p], \quad R = \operatorname{diag}[R_1, R_2, \ldots, R_p].$$

In this case, the product in (1.1) becomes

$$(1.3) \qquad \begin{bmatrix} X_1 & X_2 & \cdots & X_p \\ X_1T_1 & X_2T_2 & \cdots & X_pT_p \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ X_1T_1^{l-1} & X_2T_2^{l-1} & \cdots & X_pT_p^{l-1} \end{bmatrix} \begin{bmatrix} R_1 & 0 & \cdots & \cdot & 0 \\ 0 & R_2 & \cdots & \cdot & \cdot \\ \cdot & 0 & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & 0 & \cdot \\ \cdot & \cdot & \cdots & R_{p-1} & 0 \\ 0 & \cdot & \cdots & 0 & R_p \end{bmatrix}$$

$$\times \begin{bmatrix} X_1 & X_2 & \cdots & X_p \\ X_1T_1 & X_2T_2 & \cdots & X_pT_p \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ X_1T_1^{l-1} & X_2T_2^{l-1} & \cdots & X_pT_p^{l-1} \end{bmatrix}^*$$

In the sequel, we confine our attention to the inverse problem of representing the given Hermitian block Toeplitz matrix A in form (1.1) with an invertible left (and hence right) factor denoted by

$$\tilde{X} \doteq \operatorname{col}\,(XT^j)_{j=0}^{l-1}.$$

It is shown next that the existence of such a representation is related to solvability of a certain (block) equation with the coefficient matrix A. Moreover, this equation provides a link between a decomposition of A and similar decompositions of its singular Hermitian extensions.

THEOREM 1.2. *Given a Hermitian block Toeplitz matrix A, let there exist a representation of the form* (1.1) *with nonsingular left and right factors such that the condition*

$$TRT^* = R \tag{1.4}$$

is fulfilled. Then there exists a unique matrix Z producing a singular Hermitian extension $\hat{A}$ of A such that the equation

$$-[L_0L_1\cdots L_{l-1}]A = [ZA_{l-1}A_{l-2}\cdots A_1] \tag{1.5}$$

is solvable. Moreover, the matrix Z is given by the formula $Z = XT^lRX^*$, and hence the matrix $\hat{A}$ admits a decomposition

$$\hat{A} = \begin{bmatrix} \tilde{X} \\ XT^l \end{bmatrix} R \begin{bmatrix} \tilde{X} \\ XT^l \end{bmatrix}^* . \tag{1.6}$$

Proof. The representation (1.1) yields, for each $k = 0, 1, \ldots, l-1$,

$$\begin{bmatrix} A_k^* \\ A_{k-1}^* \\ \cdot \\ A_0 \\ \cdot \\ A_{l-k-1} \end{bmatrix} = \begin{bmatrix} X \\ XT \\ \cdot \\ \cdot \\ \cdot \\ XT^{l-1} \end{bmatrix} RT^{*k}X^*. \tag{1.7}$$

Since the matrix $\tilde{X} = \operatorname{col}\,(XT^j)_{j=0}^{l-1}$ is nonsingular, it follows (see [9]) that there exists a (unique) monic $n \times n$ matrix polynomial of degree l:

$$L(\lambda) = L_0 + \lambda L_1 + \cdots + \lambda^{l-1}L_{l-1} + \lambda^l I,$$

such that (X, T) is its (right) standard pair. That is, $\sum_{j=0}^{l} L_jXT^j = 0$, where $L_l = I$ or, what is equivalent,

$$-[L_0L_1\cdots L_{l-1}]\tilde{X} = XT^l. \tag{1.8}$$

Hence, premultiplying (1.7) by $[L_0L_1\cdots L_{l-1}]$, we obtain for $k = 0, 1, \ldots, l-1$,

$$-[L_0L_1\cdots L_{l-1}] \begin{bmatrix} A_k^* \\ A_{k-1}^* \\ \cdot \\ A_0 \\ \cdot \\ A_{l-k-1} \end{bmatrix} = XT^lRT^{*k}X^*.$$

Therefore,

$$-[L_0L_1\cdots L_{l-1}]A = [XT^lRT^* \; XT^lRT^*X^* \cdots XT^lRT^{*l-1}X^*].$$

It remains to observe that (1.4) and (1.7) yield for $k = 1, 2, \ldots, l-1$,

$$XT^lRT^{*k}X^* = XT^{l-1}RT^{*k-1}X^* = A_{l-k}, \tag{1.9}$$

to deduce the equation in (1.5) with $Z = XT^lRX^*$.

To derive (1.6), add to (1.9) the equation $XT^lRT^{*l}X^* = XT^{l-1}RT^{*l-1}X^* = XRX^* = A_0$, and observe that the system of these equations is equivalent to the decomposition (1.6). The singularity of $\hat{A}$ generated from A by the matrix Z above is verified in the following way. Rewrite (1.8) in the form

$$[L_0L_1\cdots L_{l-1}I]\begin{bmatrix}\tilde{X}\\ XT^l\end{bmatrix} = 0,$$

to obtain

$$-[L_0L_1\cdots L_{l-1}I]\hat{A} = 0.$$

The latter shows the existence of a (left) kernel of $\hat{A}$ of dimension n. Hence (see, for instance, [13, p. 95]) the Z-extension $\hat{A}$ of A is singular. □

Remark. Obviously, if (1.1) is valid, then there exists the following representation of the elements of A:

$$A_k = XT^kRX^* \qquad (k = 0, 1, \ldots, l-1). \tag{1.10}$$

Provided the condition (1.4) is fulfilled, the converse is also true. That is, (1.10) yield (1.1) (compare with Theorem 1.1).

2. The Stein equation and singular extensions. We have seen that the representation (1.1) of a Hermitian block Toeplitz matrix A is closely related to solvability of a (block) Yule–Walker-type equation. It turns out that, for a nonsingular A, this equation, with an appropriate choice of Z in (1.5), allows us to view the matrix A as a solution of a certain matrix equation.

THEOREM 2.1. *Given a nonsingular Hermitian $ln \times ln$ block Toeplitz matrix A, let the $n \times n$ matrices $L_0, L_1, \ldots, L_{l-1}$ form a solution of the equation*

$$-[L_0L_1\cdots L_{l-1}]A = [ZA_{l-1}A_{l-2}\cdots A_1] \tag{2.1}$$

with an $n \times n$ matrix Z. Denote by C_L the companion matrix in the form (0.5) *associated with the monic matrix polynomial $L(\lambda) = \sum_{j=0}^{l-1}\lambda^jL_j + \lambda^lI$. Then the homogeneous Stein equation*

$$C_LAC_L^* = A \tag{2.2}$$

holds if and only if the matrix Z produces a singular Hermitian extension of A.

Proof. Let (2.2) hold. Comparing the last block rows on both sides, we have that

$$-[L_0L_1\cdots L_{l-1}]AC_L^* = [A_{l-1}A_{l-2}\cdots A_0]. \tag{2.3}$$

Denote $-[L_0L_1\cdots L_{l-1}]A = [A_lA'_{l-1}\cdots A'_1]$. A comparison of the first $l-1$ block entries in (2.3) shows that $A'_j = A_j$ for $j = 1, 2, \ldots, l-1$, and henceforth the equation in (2.1) holds with $Z = A_l$. To check that the Hermitian Z-extension $\hat{A}$ of A is singular, we first note that a comparison of the last block entries in (2.3) implies that

$$A_lL_0^* + A_{l-1}L_1^* + \cdots + A_lL_{l-1}^* + A_0 = 0. \tag{2.4}$$

Now represent $\hat{A}$ in the form

$$\hat{A} = \begin{bmatrix}A & B^*\\ B & A_0\end{bmatrix}, \tag{2.5}$$

where $B = [ZA_{l-1} \cdots A_1]$, and observe that (2.4) yields

$$[L_0 L_1 \cdots L_{l-1} I] \begin{bmatrix} B^* \\ A_0 \end{bmatrix} = 0.$$

The latter equation, along with (2.1) rewritten in the form

$$[L_0 L_1 \cdots L_{l-1} I] \begin{bmatrix} A \\ B \end{bmatrix} = 0,$$

implies, in view of (2.5), the relation

$$[L_0 L_1 \cdots L_{l-1} I] \hat{A} = 0.$$

Thus the singularity of the Hermitian Z-extension of A is established.

Conversely, assume that the Hermitian Z-extension $\hat{A}$ of A is singular. The singularity of $\tilde{A}$ yields the existence of $n \times n$ matrices $U_0, U_1, \ldots, U_l$ such that rank $[U_0 U_1 \cdots U_l] = n$ and

$$(2.6) \qquad [U_0 U_1 \cdots U_l] \hat{A} = 0.$$

Recalling the partition (2.5) of $\hat{A}$, we deduce, in view of (2.3), the equation

$$[U_0 U_1 \cdots U_{l-1}] A + U_l [Z A_{l-1} \cdots A_1] = 0.$$

Using (2.1) and the nonsingularity of A, we have

$$[U_0 U_1 \cdots U_{l-1} U_l] = U_l [L_0 L_1 \cdots L_{l-1} I].$$

The matrix on the left is of full rank and henceforth U_l must be nonsingular. This, along with (2.6), yields the equation

$$(2.7) \qquad [L_0 L_1 \cdots L_{l-1} I] \hat{A} = 0.$$

Now observe that (2.1) with an arbitrary Z yields the (symmetric) Stein equation

$$(2.8) \qquad A - C_L A C_L^* = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \cdot & \cdot & \cdots & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & 0 \\ 0 & 0 & 0 & \cdots & 0 & W \end{bmatrix},$$

where

$$(2.9) \qquad W = Z L_0^* + \sum_{k=1}^{l} A_{l-k} L_k^* \qquad (L_l = I).$$

Indeed, it is easily seen that, in presence of (2.1), we have

$$C_L A C_L^* = \begin{bmatrix} A_1 & A_0 & \cdots & A_{l-2}^* \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ A_{l-1} & \cdot & \cdots & A_0 \\ Z & A_{l-1} & \cdots & A_1 \end{bmatrix} C_L^* = \begin{bmatrix} A_0 & A_1^* & \cdots & A_{l-1}^* \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ A_{l-2} & \cdot & \cdots & A_1^* \\ A_{l-1} & A_{l-2} & \cdots & (A_0 - W) \end{bmatrix},$$

where W is defined in (2.9). It remains to note that the condition (2.7) implies $W = 0$, and hence the equation (2.2) follows. □

Note that in the sufficient part of the theorem the assumption of nonsingularity of A is redundant. In this case the validity of (2.2) implies solvability of (2.1) with some matrix Z generating a singular Hermitian extension of A.

Remark. A slightly different relation between matrix equations in block Toeplitz matrices and the Stein equation has been indicated in [11] and [14]. In the scalar case, a connection between (2.2) and singular extensions of A can be found in [5].

The results obtained above are partially incorporated into the following theorem.

THEOREM 2.2. *Let the $ln \times ln$ block Toeplitz matrix A be nonsingular and Hermitian. Preserving the previous notation, let (X, T) stand for a (right) standard pair of the monic matrix polynomial $L(\lambda)$. Then the following statements are equivalent:*

(a) *There exists a representation*

$$A = \begin{bmatrix} X \\ XT \\ \cdot \\ \cdot \\ \cdot \\ XT^{l-1} \end{bmatrix} R \begin{bmatrix} X \\ XT \\ \cdot \\ \cdot \\ \cdot \\ XT^{l-1} \end{bmatrix}^* \tag{2.10}$$

in which $TRT^* = R$.

(b) *The Hermitian extension $\hat{A}$ of A produced by the matrix $Z = XT^lRX^*$ is singular and admits a representation*

$$\hat{A} = \begin{bmatrix} \tilde{X} \\ XT^l \end{bmatrix} R \begin{bmatrix} \tilde{X} \\ XT^l \end{bmatrix}^*, \tag{2.11}$$

where $\tilde{X} = \operatorname{col}(XT^j)_{j=0}^{l-1}$.

(c) *The equation*

$$[L_0L_1 \cdots L_{l-1}I]\hat{A} = 0 \tag{2.12}$$

holds.

(d) *The homogeneous Stein equation*

$$C_LAC_L^* = A \tag{2.13}$$

is valid.

Proof. The implication **1** → **2** is a consequence of Theorem 1.2. Let the statement in **2** be valid. As shown in the proof of Theorem 2.1, the singularity of $\hat{A}$ yields the equation (2.12) and subsequently, (2.13). Hence the implication **2** → **3** → **4** is verified. The implication **4** → **1** is shown in the following way. Define the matrix

$$R = \tilde{X}^{-1}A\tilde{X}^{*-1}.$$

Obviously, the representation (2.10) holds. Now substitute into (2.13) the easily verified relation

$$C_L = \tilde{X}T\tilde{X}^{-1}$$

and use the representation (2.10) to deduce the required relation $TRT^* = R$. □

Note that for a nonsingular A, (2.13) yields the nonsingularity of the companion matrix C_L. In particular, this implies $\det L_0 \neq 0$ and hence the monic matrix polynomial

$L(\lambda)$ has no zero eigenvalues. Furthermore, it is well known that any $p \times p$ nonsingular Hermitian matrix R is congruent (see, e.g., [13, p. 184]) to the matrix

$$D = \begin{bmatrix} I_s & 0 \\ 0 & -I_{p-s} \end{bmatrix},$$

where s denotes the number of positive eigenvalues of R counting multiplicities. Hence the decomposition (2.10) can be written in the form

$$A = [\operatorname{col}(\underline{X}\underline{T}^j)_{j=0}^{l-1}]D[\operatorname{col}(\underline{X}\underline{T}^j)_{j=0}^{l-1}]^*,$$

where $WDW^* = R$ with a nonsingular matrix W and $\underline{X} = XW$, $\underline{T} = W^{-1}TW$. Note that, along with (X, T), the pair of matrices $(\underline{X}, \underline{T})$ is a right standard pair of the monic matrix polynomial $L(\lambda)$ associated with $A = [A_{i-j}]_{i,j=0}^{l-1}$ and its singular extension. It is also easily observed that $\underline{T}^*$ is D-unitary.

Thus one can view (2.10) as a representation of the Hermitian block Toeplitz matrix A in a form of a *generalized Gram matrix*:

$$A = [\langle \underline{T}^{*j}G, \underline{T}^{*i}G\rangle]_{i,j=0}^{l-1},$$

in which $G = \underline{X}^*$ and the indefinite (block) inner product of one-block column matrices $\langle E, F\rangle \doteq (E, DF) \doteq F^*DE$ is used. Clearly, in the scalar positive definite case this reduces to a representation of the Toeplitz matrix A as an ordinary Gram matrix.

3. The main theorem. Let A be a nonsingular Hermitian block Toeplitz matrix satisfying the equation

$$C_L A C_L^* = A$$

with a companion matrix C_L associated with a monic matrix polynomial $L(\lambda)$. Note that the equation above shows that the matrix C_L^* is A-unitary.

Furthermore, as defined in [10], two pairs of matrices (B_1, H_1) and (B_2, H_2) are said to be *unitarily similar* if there exists a nonsingular matrix S such that

$$B_1 = S^{-1}B_2S, \qquad H_1 = S^*H_2S.$$

Thus the relations

$$C_L = \tilde{X}T\tilde{X}^{-1}, \qquad A = \tilde{X}R\tilde{X}^*$$

from above (with the first rewritten in the complex conjugate form) mean that the pairs (C_L^*, A) and (T^*, R) are unitarily similar with the transition matrix $S = \tilde{X}^*$. In particular, T^* is quasi-unitary with respect to the indefinite metric generated by R, and hence the spectrum of T is symmetric relative to the unit circle. Recall that (X, T) is *any* (right) standard pair of the monic matrix polynomial $L(\lambda)$. It is the goal of this section to determine a standard pair for $L(\lambda)$ such that the decomposition of the given block Toeplitz matrix is as simple as possible. As expected, the desirable pair is related to a (specific) Jordan pair.

Let (X^0, J) stand for a Jordan pair of $L(\lambda)$, in which the blocks of the Jordan matrix J are arranged as follows:

$$J = \operatorname{diag}[J_1, J_2, \ldots, J_\alpha, J_{\alpha+1}, \ldots, J_{\alpha+2\beta}], \tag{3.1}$$

where each J_j is a Jordan block; $J_1, \ldots, J_\alpha$ have their associated eigenvalue on the unit circle; the eigenvalues of $J_{\alpha+1}, J_{\alpha+3}, \ldots, J_{\alpha+2\beta-1}$ are outside the unit circle; and the eigenvalues of $J_{\alpha+2j}$ ($j = 1, 2, \ldots, \beta$) are obtained from that of $J_{\alpha+2j-1}$ by inversion in the unit circle.

It is shown in [10, p. 79] that the pair (T^*, R) (as well as (C_L^*, A)) is unitarily similar to the pair (J, Q_t), called a *canonical pair*, where

$$(3.2)\quad Q_\varepsilon = \operatorname{diag}\left[\varepsilon_1 Q_1, \varepsilon_2 Q_2, \ldots, \varepsilon_\alpha Q_\alpha, \begin{bmatrix} 0 & Q_{\alpha+1} \\ Q_{\alpha+1}^* & 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 & Q_{\alpha+\beta} \\ Q_{\alpha+\beta}^* & 0 \end{bmatrix}\right].$$

The size of Q_j equals that of J_j for $j = 1, 2, \ldots, \alpha$ and equals that of $J_{\alpha+2(j-\alpha)}$ for $j > \alpha$. The ordered set $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_\alpha$ of signs ± 1 is uniquely determined by (T^*, R) up to permutation of signs corresponding to equal blocks Q_j. Clearly, the matrix J is Q_ε-unitary.

The structure of the matrices Q_j ($j = 1, 2, \ldots, \alpha + \beta$) in (3.2) can be described as follows (cf. [10, p. 79]): For a *unimodular eigenvalue* λ:

$$(3.3)\qquad Q_j = S_j^* P_j S_j \qquad (j = 1, 2, \ldots, \alpha),$$

where P_j is an $r_j \times r_j$ sip matrix $[\delta_{i,r_j+1-k}]_{i,k=1}^{r_j}$ ($\delta_{i,k}$ is the symbol of Kronecker),

$$S_j = \operatorname{diag}[1, i^{-1}, i^{-2}, \cdots] M_j \operatorname{diag}[1, (2\lambda)^{-1}, (2\lambda)^{-2}, \cdots]$$

and

$$M_j = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & \cdot & 0 \\ 0 & 1 & -1 & 1 & -1 & \cdots & \cdot & (-1)^{r_j} \\ 0 & 0 & 1 & -2 & 3 & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & 1 & -3 & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 1 & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & -r_j+2 \\ 0 & 0 & \cdot & \cdot & \cdot & \cdots & 0 & 1 \end{bmatrix}.$$

Note that the columns of M_j are made up of the signed binomial coefficients.

For a *nonunimodular eigenvalue* λ it is found that, preserving the previous notation,

$$(3.4)\qquad Q_j = S_{1j}^* P_j S_{2j} \qquad (j = \alpha + 1, \alpha + 2, \ldots, \alpha + \beta),$$

where

$$S_{1j} = \operatorname{diag}[1, \{0.5i(1+\lambda)\}^{-1}, \{0.5i(1+\lambda)\}^{-2}, \cdots] M_j \\ \times \operatorname{diag}[1, (1+\lambda)^{-1}, (1+\lambda)^{-2}, \cdots],$$

and S_{2j} is obtained from S_{1j} on replacing λ by $\bar{\lambda}^{-1}$.

The unitary similarity of the pairs (T^*, R) and (J, Q_ε) means that there exists a nonsingular matrix S such that

$$(3.5)\qquad T^* = S^{-1} J S, \qquad R = S^* Q_\varepsilon S.$$

Hence $T = S^* J^* S^{*-1}$ and the decomposition $A = \tilde{X} R \tilde{X}^*$ above can be written in the form

$$(3.6)\qquad A = \begin{bmatrix} \hat{X} \\ \hat{X} J^* \\ \cdot \\ \cdot \\ \hat{X} J^{*l-1} \end{bmatrix} Q_\varepsilon \begin{bmatrix} \hat{X} \\ \hat{X} J^* \\ \cdot \\ \cdot \\ \hat{X} J^{*l-1} \end{bmatrix}^*,$$

where the matrix $\hat{X} = XS^*$ is the first component in a Jordan pair for $L(\lambda)$. Note that the matrices $\hat{X}$ and X^0 do not necessarily coincide.

Now combine the neighboring Jordan blocks in (3.1) associated with nonunimodular eigenvalues and write

$$J = \operatorname{diag}[\hat{J}_1, \hat{J}_2, \dots, \hat{J}_\alpha, \hat{J}_{\alpha+1}, \dots, \hat{J}_{\alpha+\beta}], \tag{3.7}$$

where $\hat{J}_j = J_j$ for $j = 1, 2, \dots, \alpha$ and $\hat{J}_j = \operatorname{diag}[J_j, J_{j+1}]$ when $j = \alpha + 1, \alpha + 3, \dots, \alpha + 2\beta - 1$. Decompose $\hat{X}$ in accordance with that of J:

$$\hat{X} = [X_1 \cdots X_\alpha X_{\alpha+1} \cdots X_{\alpha+\beta}]. \tag{3.8}$$

Denoting $p = \alpha + \beta$, we summarize the previous discussion in the following theorem.

THEOREM 3.1. *Let A denote a nonsingular Hermitian block Toeplitz matrix*

$$A = [A_{i-j}]_{i,j=1}^{l} \qquad (A_k \in \mathbf{C}^{n\times n}, A_{-k} = A_k^*, k = 0, 1, \dots, l = 1).$$

Then to every $n \times n$ matrix Z such that

$$\operatorname{rank}\begin{bmatrix} A_0 & A_1^* & \cdots & A_{l-1}^* & Z^* \\ A_1 & A_0 & \cdots & \cdot & A_{l-1}^* \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ A_{l-1} & \cdot & \cdots & \cdot & A_1^* \\ Z & A_{l-1} & \cdots & A_1 & A_0 \end{bmatrix} = \operatorname{rank} A, \tag{3.9}$$

there corresponds a unique set of integers $r_1, r_2, \dots, r_p$ ($\sum r_k = ln$) such that the representation

$$A = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \\ X_1\hat{J}_1^* & X_2\hat{J}_2^* & \cdots & X_p\hat{J}_p^* \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ X_1\hat{J}_1^{*l-1} & X_2\hat{J}_2^{*l-1} & \cdots & X_p\hat{J}_p^{*l-1} \end{bmatrix}$$

$$\times \begin{bmatrix} R_1 & 0 & \cdots & \cdot & 0 \\ 0 & R_2 & \cdots & \cdot & \cdot \\ \cdot & 0 & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & 0 & \cdot \\ \cdot & \cdot & \cdots & R_{p-1} & 0 \\ 0 & \cdot & \cdots & 0 & R_p \end{bmatrix} \begin{bmatrix} X_1 & X_2 & \cdots & X_p \\ X_1\hat{J}_1^* & X_2\hat{J}_2^* & \cdots & X_p\hat{J}_p^* \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ X_1\hat{J}_1^{*l-1} & X_2\hat{J}_2^{*l-1} & \cdots & X_p\hat{J}_p^{*l-1} \end{bmatrix}^* \tag{3.10}$$

with the following properties is valid:

(a) *The matrices R_j are $r_j \times r_j$ nonsingular Hermitian and are given by the formulas*

$$R_j = \varepsilon_j Q_j \quad (j = 1, 2, \dots, \alpha), \quad R_j = \begin{bmatrix} 0 & Q_j \\ Q_j^* & 0 \end{bmatrix} \quad (j = \alpha + 1, \alpha + 2, \dots, p), \tag{3.11}$$

where the integer α is defined above and the matrices Q_j are given in (3.3) *and* (3.4).

(b) *The $r_j \times r_j$ matrices $\hat{J}_j$ are Jordan blocks when associated with unimodular eigenvalues and are direct sums of two Jordan blocks corresponding to nonunimodular eigenvalues mutually symmetric relative to the unit circle.*

(c) *The matrices X_j are $n \times r_j$ and such that the matrices*

$$\hat{X} = [X_1 X_2 \cdots X_p], \qquad J = \operatorname{diag}[\hat{J}_1, \hat{J}_2, \dots, \hat{J}_p]$$

constitute a (right) standard pair of a monic matrix polynomial. In particular, the left and right factors in (3.10) *are nonsingular.*

(d) *The* $(l+1)n \times (l+1)n$ *matrix* $\hat{A}$ *in* (3.9) *admits representation*

$$\hat{A} = \begin{bmatrix} \tilde{X} \\ X\hat{J}^{*l} \end{bmatrix} R \begin{bmatrix} \tilde{X} \\ X\hat{J}^{*l} \end{bmatrix}^*, \tag{3.12}$$

where $\tilde{X}$ *and* R *denote the left and the middle factor in* (3.10), *respectively. In particular, it holds that*

$$Z = \sum_{j=1}^{p} X_j \hat{J}_j^{*l} R_j X_j^*. \tag{3.13}$$

Conversely, for any representation (3.10) *with the indicated properties, the matrix* Z *in* (3.13) *defines a (singular) Hermitian extension* $\hat{A}$ *of* A *such that* (3.12) *holds.*

Proof. Let a matrix Z satisfying (3.9) be fixed. By Theorem 2.1, (2.1) generates a unique monic matrix polynomial $L(\lambda) = \sum_{j=0}^{l} \lambda^j L_j$ such that

$$C_L A C_L^* = A.$$

Furthermore, it follows from Theorem 2.2 that, for any (right) standard pair (X, T) of $L(\lambda)$, there exists a Hermitian matrix R such that

$$A = \tilde{X} R \tilde{X}^*, \qquad TRT^* = R,$$

where $\tilde{X} = \operatorname{col}(XT^j)_{j=0}^{l-1}$.

As pointed out above, a passage from the pair (T^*, R) to its (unitarily similar) canonical form (J, Q_ε), where J is arranged as in (3.1) and Q_ε is defined in (3.2), allows one to derive the factorization (3.6) in which $\hat{X}$ is the first component in the (right) pair $(\hat{X}, J^*)$ for $L(\lambda)$. In view of (3.7) and (3.8), the representation (3.10) now follows from (3.6) on using the substitution (3.11). Let r_j denote the size of the matrix $\hat{J}_j$ in (3.7). Then the matrix X_j in (3.8) is $n \times r_j$. Furthermore, for every j the sizes of $\hat{J}_j$ and Q_j coincide and hence the matrix R_j is $r_j \times r_j$, as stated. Note that the set of indices $\{r_j\}$ is uniquely defined by the monic matrix polynomial $L(\lambda)$. The observation that the decomposition (3.12) follows from Theorem 2.2 completes the proof of the first part of the theorem.

Conversely, let (3.10) be valid. To derive the existence of a representation (3.12) for $\hat{A}$, we first observe that (3.10) in the compact form $A = \tilde{X} R \tilde{X}^*$ with $J^*RJ = R$ yields

$$C_L A C_L^* = C_L \tilde{X} R \tilde{X}^* C_L^* = \tilde{X} J^* R J \tilde{X}^* = \tilde{X} R \tilde{X}^* = A,$$

where $L(\lambda)$ is the monic matrix polynomial having a (right) standard pair $(\hat{X}, J^*)$. According to Theorem 2.2, in which the matrices X, T are replaced by $\hat{X}$, J^*, respectively, the required decomposition (3.12) holds provided that the (singular) extension $\hat{A}$ is generated by the matrix $\hat{X} J^{*l} R \hat{X}$. The latter coincides with the matrix Z in (3.13). This completes the proof. □

Observe that Theorem 0.2 announced in the Introduction follows immediately from Theorem 3.1.

Remark. The decomposition $A = \tilde{X} R \tilde{X}^*$ with the indicated properties can be rewritten in the following equivalent form. Let a singular extension of A be found and the corresponding monic matrix polynomial $L(\lambda)$ be computed by formula (2.12). Denote by J a Jordan matrix associated with $L(\lambda)$ and fix a matrix X^0 such that (X^0, J^*) is a (right) standard pair of $L(\lambda)$. Note that all matrices which can be substituted instead of

X^0 in (3.10) are given by the formula $X = X^0 M$, where M is any nonsingular matrix satisfying the condition

$$(3.14) \qquad MJ^* = J^*M \qquad (\det M \neq 0).$$

Hence (3.10) can be written in the (compact) form

$$(3.15) \qquad A = \begin{bmatrix} X^0 \\ X^0 J^* \\ \cdot \\ \cdot \\ \cdot \\ X^0 J^{*l-1} \end{bmatrix} MRM^* \begin{bmatrix} X^0 \\ X^0 J^* \\ \cdot \\ \cdot \\ \cdot \\ X^0 J^{*l-1} \end{bmatrix}^* .$$

Note that M is a direct sum of matrices M_j, each of the size of $\hat{J}_j$, satisfying the equation $M_j \hat{J}_j^* = \hat{J}_j^* M_j$ for each j.

Formula (3.15) is helpful in computing decompositions of the form (3.10). Indeed, provided a pair (X^0, J^*) is found, it remains to determine M satisfying (3.14) in order to obtain a decomposition of the required type. □

A consequence of Theorem 3.1 regarding Hermitian positive definite block Toeplitz matrices is presented next.

THEOREM 3.2. *A Hermitian block Toeplitz matrix* $A = [A_{i-j}]_{i,j=1}^{l}$ *is positive definite if and only if the* $n \times n$ *entries of* A *admit a representation*

$$(3.16) \qquad A_k = \sum_{j=1}^{ln} \rho_j \mu_j^k X_j X_j^* \qquad (k = 0, \pm 1, \dots, \pm(l-1)),$$

in which X_j *are* $n \times 1$ *vectors*, $\rho_j > 0$, *and the numbers* μ_j *are unimodular, and such that the matrix*

$$\tilde{X} \doteq \begin{bmatrix} X_1 & X_2 & \cdots & X_p \\ \mu_1 X_1 & \mu_2 X_2 & \cdots & \mu_p X_p \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mu_1^{l-1} X_1 & \mu_2^{l-1} X_2 & \cdots & \mu_p^{l-1} X_p \end{bmatrix} \qquad (p = ln)$$

is nonsingular.

Proof. Let the matrix A be positive definite. Then (see [6]) there exists a singular Hermitian Z-extension $\hat{A}$ of A and, subsequently, Theorem 3.1 asserts the existence of a representation of the form (3.10) in which the matrices R_j are necessarily positive definite. Hence it follows from (3.11) that $p = \alpha$ and Q_j must be definite Hermitian matrices. The sip matrices of order $j > 1$ fail to be definite and therefore, in view of (3.3), all Q_j turn out to be 1×1 matrices. Thus $p = ln$ and R_j in (3.10) are positive numbers denoted by ρ_j $(j = 1, 2, \dots, ln)$.

The dimension of Q_j is equal to that of $\hat{J}_j = J_j$ and, consequently, the Jordan form J is a diagonal matrix with unimodular eigenvalues:

$$J = \operatorname{diag}[\bar{\mu}_1, \bar{\mu}_2, \dots, \bar{\mu}_{ln}].$$

Thus the representation

$$(3.17) \qquad A = \tilde{X} R \tilde{X}^*,$$

where

$$R = \operatorname{diag}\,[\rho_1, \rho_2, \ldots, \rho_p], \quad \tilde{X} = \operatorname{col}\,(XJ^{*j})_{j=0}^{l-1}, \quad X = [X_1 X_2 \cdots X_p] \quad (p = ln)$$

is established. Obviously, it is equivalent to (3.16). Conversely, a representation of the form (3.16) (or, what is equivalent, (3.17)) with $\rho_j > 0$, yields the positive definiteness of A. □

Note that the representations (3.16) can be viewed as decompositions of the entries of a block Toeplitz into sums of matrices of rank one. Furthermore, the vectors X_j in Theorem 3.2 are (ordinary, not generalized) eigenvectors of the monic matrix polynomial associated with the specific singular extension of the given matrix. The numbers μ_j are not necessarily distinct, but enjoy the property det $\tilde{X} \neq 0$, which follows from the fact that $(\hat{X}, J^*)$ is a pair of a monic matrix polynomial.

COROLLARY 3.3 (see [1]–[3]). *A Hermitian Toeplitz matrix* $A = [a_{i-j}]$ *of order* l *is positive definite if and only if its entries allow a representation*

$$a_k = \sum_{j=1}^{l} \rho_j \mu_j^k \qquad (k = 0, \pm 1, \ldots, \pm(l-1)),$$

in which $\rho_j > 0$, $|\mu_j| = 1$ $(j = 1, 2, \ldots, l)$ *and all* μ_j *are distinct.*

Proceeding to the general case of a (scalar) nonsingular Hermitian Toeplitz matrix, we first recall that the matrix J in (3.7) coincides with the Jordan form of the (scalar) companion matrix C_L and, since the latter is nonderogatory, the Jordan blocks in J correspond to distinct eigenvalues. (Note that, for a similar reason, the number of cells in the Jordan form of an $n \times n$ block companion matrix corresponding to equal eigenvalues cannot exceed n.) Furthermore, recalling the representation (3.15), we note that in the scalar case, the entries of the matrix X^0 can be chosen as 1's for each simple root of $L(\lambda)$ and the vector $[0 \cdots 0\ 1]$ with $k - 1$ zeros for a root of multiplicity k. These observations lead to the following consequence of Theorem 3.1 concerning nonsingular (scalar) Hermitian Toeplitz matrices and given here in a form based on the remark to Theorem 3.1.

COROLLARY 3.4. *Let* $A = [a_{i-j}]$ *denote a nonsingular Hermitian Toeplitz matrix of order* l. *Then, to every complex number* z *generating a singular Hermitian extension of* A, *there corresponds a unique set of integers* $r_1, r_2, \ldots, r_p$ $(\sum_{j=1}^{p} r_j = l)$ *such that the entries of* A *admit representation*

$$a_k = \sum_{j=1}^{p} X_{0j} \hat{J}_j^{*k} M_j R_j M_j^* X_{0j}^* \qquad (k = 0, \pm 1, \ldots, \pm(l-1)), \tag{3.18}$$

in which

(a) R_j *are* $r_j \times r_j$ *nonsingular Hermitian matrices computed by* (3.11), (3.3), (3.4).

(b) $\hat{J}_j$ *are* $r_j \times r_j$ *matrices having mutually disjoint spectra symmetric* (*with respect to the unit circle*). *They are Jordan blocks when associated with unimodular eigenvalues and are direct sums of two Jordan blocks corresponding to mutually symmetric eigenvalues.*

(c) $X_{0j} = [0 \cdots 0\ 1]$ (*with* $r_j - 1$ *zeros*).

(d) *the nonsingular matrices* M_j *obey the equation* $M_j \hat{J}_j^* = \hat{J}_j^* M_j$.

Furthermore,

$$z = \sum_{j=1}^{l} X_{0j} \hat{J}_j^{*l} M_j R_j M_j^* X_{0j}^*. \tag{3.19}$$

Conversely, to every representation (3.18) *with the indicated properties, the scalar* z *in* (3.19) *defines a* (*singular*) *Hermitian extension of* A.

Note that the existence of (infinitely many) Hermitian extensions of a Hermitian Toeplitz matrix is established in [12].

4. Examples. Several illustrative examples of Hermitian (scalar and block) Toeplitz matrices and their decompositions of the indicated type are presented in this section.

Example 1. Consider the nonsingular indefinite Hermitian Toeplitz matrix

$$A = \begin{bmatrix} 1 & i & 1 & -i \\ -i & 1 & i & 1 \\ 1 & -i & 1 & i \\ i & 1 & -i & 1 \end{bmatrix},$$

and observe that the number 1 generates a singular Hermitian extension $\hat{A}$ of A such that the equation (see (2.12))

$$[-1\ 0\ 0\ 0\ 1]\hat{A} = 0$$

is valid. This extension produces a monic polynomial

$$L(\lambda) = \lambda^4 - 1 = (\lambda^2 + 1)(\lambda^2 - 1)$$

with roots ±1, $\pm i$. Hence the pair of matrices

$$X^0 = [1\ 1\ 1\ 1], \qquad J^* = \operatorname{diag}[1, -1, -i, i]$$

forms a (right) standard pair for $L(\lambda)$. As shown above, there exists a factorization of A of the form $A = \tilde{X}R\tilde{X}^*$, in which

$$\tilde{X} = \operatorname{col}(XJ^{*j})_{j=0}^{3}, \quad R = \operatorname{diag}[\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4] \quad (\varepsilon_j = \pm 1),$$

and the matrices X, J^* constitute a (right) standard pair of $L(\lambda)$. To find X, we recall the remark of the preceding section and write $X = X^0 M$. Hence the decomposition required becomes, as in (3.15) with the left factor,

$$\tilde{X}^0 = (X^0, J^{*j})_{j=0}^{3} \doteq \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -i & i \\ 1 & 1 & -1 & 1 \\ 1 & -1 & i & -i \end{bmatrix},$$

which turns out to be, in this case, a usual Vandermonde matrix. Due to the fact that the Jordan matrix is diagonal, so is the matrix M above. The condition $MJ^* = J^*M$ provides no information in this case and henceforth $M = \operatorname{diag}[m_1, m_2, m_3, m_4]$ with *a priori* no restrictions on the numbers. The Vandermonde matrix $V = \tilde{X}^0$ constructed from numbers of modulus 1 is easily invertible (in this case, $V^{-1} = 0.25V^*$), and hence the equation

$$A = \tilde{X}^0 \operatorname{diag}[\varepsilon_1|m_1|^2, \varepsilon_2|m_2|^2, \varepsilon_3|m_3|^2, \varepsilon_4|m_4|^2]\tilde{X}^{0*}$$

becomes

$$V^*AV = 16 \operatorname{diag}[\varepsilon_1|m_1|^2, \varepsilon_2|m_2|^2, \varepsilon_3|m_3|^2, \varepsilon_4|m_4|^2].$$

The latter gives

$$\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = 1, \qquad \varepsilon_4 = -1, \ |m_1|^2 = |m_2|^2 = |m_3|^2 = |m_4|^2 = 0.5.$$

Choosing $m_1 = m_2 = m_3 = m_4 = 1/\sqrt{2}$, we derive the following representation of the given matrix A

$$A = \tilde{X} \operatorname{diag}[1, 1, 1, -1]\tilde{X}^*,$$

in which $X = X^0M = (1/\sqrt{2})\,[1\ 1\ 1\ 1]$. In other words, there exists the following representation of the elements a_k of A:

$$a_k = \sum_{j=1}^{4} \rho_j \mu_j^k \qquad (k = 0, 1, 2, 3),$$

where $\rho_1 = \rho_2 = \rho_3 = 1$, $\rho_4 = -1$ and $\mu_1 = 1$, $\mu_2 = -1$, $\mu_3 = i$, $\mu_4 = -i$.

Note that the decomposition above shows that the matrix A has one negative and three positive eigenvalues.

In the preceding example, the roots of the polynomial $L(\lambda)$ are distinct and unimodular. This implies, in particular, that the matrix $\tilde{X}$ is a usual Vandermonde matrix and the matrix R is diagonal. The example below deals with the case when the corresponding polynomial has roots outside the unit circle.

Example 2. Let the real symmetric indefinite Toeplitz matrix

$$A = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}$$

be given. Since A and its leading principal submatrix of order 2 are nonsingular, it follows (cf. [12, p. 99]) that there are at most two different real symmetric singular extensions of A produced in this case by the numbers 3 and -3. Denote the corresponding extensions by $\hat{A}_1$, $\hat{A}_2$ and observe that the equations

$$[1\ {-2}\ {-2}\ 1]\hat{A}_1 = 0, \qquad [-1\ {-2}\ 2\ 1]\hat{A}_2 = 0$$

give rise to monic polynomials

$$L_1(\lambda) = \lambda^3 - 2\lambda^2 - 2\lambda + 1 = (\lambda + 1)(\lambda^2 - 3\lambda + 1),$$
$$L_2(\lambda) = \lambda^3 + 2\lambda^2 - 2\lambda - 1 = (\lambda - 1)(\lambda^2 + 3\lambda + 1).$$

The roots -1, $0.5(3 \pm \sqrt{5})$ and 1, $0.5(-3 \pm \sqrt{5})$ of the polynomials $L_1(\lambda)$ and $L_2(\lambda)$, respectively, are distinct, but not all lie on the unit circle. Obviously, the nonunimodular roots are symmetric with respect to the unit circle, as required.

Coupling the symmetric roots, we obtain pairs (X_1, J_1^*) and (X_2, J_2^*) for the polynomials $L_1(\lambda)$, $L_2(\lambda)$, respectively, in which

$$X_1^0 = X_2^0 = [1\ 1\ 1], \quad J_1 = \operatorname{diag}[-1, \alpha, \alpha^{-1}], \quad J_2 = \operatorname{diag}[1, \beta, \beta^{-1}],$$

where $\alpha = 0.5(3 + \sqrt{5})$, $\beta = 0.5(-3 + \sqrt{5})$.

As seen from the formulas (3.13), (3.3), and (3.4), and due to the disjoint spectra of J_1, J_2, the middle factors R_1 and R_2 in the representations of A have to be block diagonal with 1×1 blocks corresponding to unimodular roots and 2×2 blocks associated with the pairs of roots mutually symmetric with respect to the unit circle. Proceeding as in the previous example, we derive the following factorizations of the matrix A:

$$A = \begin{bmatrix} X_1 \\ X_1J_1 \\ X_1J_1^2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_1J_1 \\ X_1J_1^2 \end{bmatrix}^*$$

and

$$A = \begin{bmatrix} X_2 \\ X_2 J_2 \\ X_2 J_2^2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_2 \\ X_2 J_2 \\ X_2 J_2^2 \end{bmatrix}^* ,$$

in which

$$X_1 = X_2 = X_1^0 M = [m_1, m_2, m_3]$$

and the numbers m_j satisfy the conditions: $|m_1|^2 = 0.6$, $|m_2|^2 = |m_3|^2 = 0.2$. Observe that both factorizations have the same middle factor and that there are no other factorizations of the indicated form.

The case of a Toeplitz matrix generating a monic polynomial with multiple (unimodular) roots is treated in the next example.

Example 3. Consider the nonsingular indefinite Toeplitz matrix

$$A = \begin{bmatrix} 1 & i & 1+2i \\ -i & 1 & i \\ 1-2i & -i & 1 \end{bmatrix}$$

and its singular Hermitian extension $\hat{A}$ produced by the number $-3i$. It is easily checked that in view of the equation

$$[1 \ -1 \ -1 \ 1]\hat{A} = 0$$

the singular extension chosen generates a monic polynomial

$$L(\lambda) = \lambda^3 - \lambda^2 - \lambda + 1 = (\lambda + 1)(\lambda - 1)^2$$

with the Jordan matrix

$$J = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

The pair (X^0, J^*), where $X^0 = [1 \ 0 \ 1]$, constitutes a (right) standard pair for $L(\lambda)$. Compute

$$\tilde{X}^0 = \operatorname{col}\,(X^0 J^{*j})_{j=0}^2 = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix}.$$

The factor M in (3.15) satisfying the condition $MJ^* = J^*M$ is a complex matrix of the form

$$M = \begin{bmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & \gamma & \beta \end{bmatrix}$$

while the matrices $R_1 = \varepsilon_1 Q_1$, $R_2 = \varepsilon_2 Q_2$ corresponding to the simple root -1 and the double root 1, respectively, are given by

$$Q_1 = 1, \qquad Q_2 = \frac{i}{2}\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

A factorization of the matrix of the required form is obtained, in particular, when $\varepsilon_1 = 1$, $\varepsilon_2 = -1$, $\alpha = 1/\sqrt{2}$, $\beta = \sqrt{2}$, and $\gamma = i/2\sqrt{2}$, and is given below:

$$A = \begin{bmatrix} X \\ XJ^* \\ XJ^{*2} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & \frac{i}{2} \\ 0 & \frac{-i}{2} & 0 \end{bmatrix} \begin{bmatrix} X \\ XJ^* \\ XJ^{*2} \end{bmatrix}^*, \quad \text{where}$$

$$X = \left[\frac{1}{\sqrt{2}}, \sqrt{2}; \frac{i}{2\sqrt{2}}\right].$$

The next example is concerned with a positive definite block Toeplitz matrix.

Example 4. The real symmetric block Toeplitz matrix

$$A = \begin{bmatrix} A_0 & A_1^T \\ A_1 & A_0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 2 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

is positive definite and its singular extension $\hat{A}$ by the matrix $Z = 2I$ generates the monic matrix $L(\lambda) = (\lambda^2 - 1)I$ derived from the equation

$$\begin{bmatrix} -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix} \hat{A} = 0.$$

A Jordan pair of $L(\lambda)$ is given by the matrices

$$X^0 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \qquad J = J^* = \operatorname{diag}[1, 1, -1, -1].$$

Hence the factorization

$$A = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1.5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & .5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}^*,$$

which, by an appropriate choice of the (diagonal) matrix $M = \operatorname{diag}[m_1, m_2, m_3, m_4]$, can be written in the required form (namely, with an identity matrix as its middle factor). Obviously, this can be achieved by setting $m_1 = \sqrt{1.5}$, $m_2 = m_4 = 1$, $m_3 = \sqrt{0.5}$. With this choice, the following representations of the elements of A are valid (compare with (3.16)):

$$A_0 = \sum_{j=1}^{4} \rho_j X_j X_j^*, \qquad A_1 = \sum_{j=1}^{4} \rho_j \mu_j X_j X_j^*,$$

where $\rho_j = 1$, $\mu_1 = \mu_2 = 1$, $\mu_3 = \mu_4 = -1$, and

$$X_1 = \sqrt{1.5}[1\ 0]^T, \quad X_3 = \sqrt{0.5}[1\ 0]^T, \quad X_2 = X_4 = [0\ 1]^T.$$

We conclude with an example of an indefinite Hermitian block Toeplitz matrix.

Example 5. Consider the following block Toeplitz matrix with 2 × 2 blocks:

$$A = \begin{bmatrix} A_0 & A_1^T \\ A_1 & A_0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

The matrix A is nonsingular real symmetric and it is easily verified that the equation

$$[L_0\, L_1\, I] \begin{bmatrix} A_0 & A_1^T & Z^T \\ A_1 & A_0 & A_1^T \\ Z & A_1 & A_0 \end{bmatrix} = 0$$

is valid for

$$Z = \begin{bmatrix} 7 & 0 \\ 0 & -2 \end{bmatrix}, \quad L_0 = I, \quad L_1 = \begin{bmatrix} -4 & 0 \\ 0 & 0 \end{bmatrix}.$$

The eigenvalues of the 2 × 2 monic matrix polynomial generated by the Z-extension of A,

$$L(\lambda) = \begin{bmatrix} \lambda^2 - 4\lambda + 1 & 0 \\ 0 & \lambda^2 + 1 \end{bmatrix},$$

are $\mu_{1,2} = \pm i$, $\mu_{3,4} = 2 \pm \sqrt{3}$. Hence there are two distinct unimodular roots and a pair of mutually symmetric roots. The matrices

$$X^0 = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \qquad J^* = \operatorname{diag}\,[i, -i, \alpha, \alpha^{-1}],$$

where $\alpha = 2 + \sqrt{3}$, constitute a (right) pair of $L(\lambda)$ and therefore the following factorization of the matrix A,

$$A = \tilde{X}R\tilde{X}^* = \begin{bmatrix} 0 & 0 & m_3 & m_4 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & \alpha m_3 & \alpha^{-1}m_4 \\ i & -i & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & m_3 & m_4 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & \alpha m_3 & \alpha^{-1}m_4 \\ i & -i & 0 & 0 \end{bmatrix}^*$$

with numbers m_3, m_4 satisfying $|m_3|^2 = |m_4|^2 = 0.5$, is valid.

A different singular (symmetric) extension of A produces another factorization of the matrix. For instance, if

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix},$$

then the Z-extension of A generates the monic matrix polynomial

$$L(\lambda) = \begin{bmatrix} \lambda^2 - 1 & 0 \\ 0 & \lambda^2 - 1 \end{bmatrix}$$

with a Jordan pair given in the preceding example. Hence the factorization

$$A = \begin{bmatrix} m_1 & 0 & m_3 & 0 \\ 0 & 1 & 0 & 1 \\ m_1 & 0 & -m_3 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} m_1 & 0 & m_3 & 0 \\ 0 & 1 & 0 & 1 \\ m_1 & 0 & -m_3 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}^*,$$

in which $|m_1|^2 = 1.5$ and $|m_3|^2 = 0.5$.

Acknowledgments. The author thanks Dr. R. Janssen for encouraging this work and Prof. P. Lancaster for useful suggestions.

REFERENCES

[1] N. I. Achiezer and M. G. Krein, *The L-problem of moments*, in Some Questions in the Theory of Moments, Tr. nsl. Math. Monographs 2, American Mathematical Society, Providence, RI, 1962.

[2] C. Caratheodory, *Ueber den Variabilitaetsbereich der Koeffizienten von Potenzreihen*, die degebene Werte nicht annehmen, Math. Annalen, G4 (1907), pp. 93–115.

[3] ———, *Ueber den Variabilitaetsbereich der Fourierschen Konstanten von positiven Funktionen*, Rediconti del Circolo Matem. di Palermo, 32 (1911), pp. 93–107.

[4] P. Delsarte, Y. Genin, and Y. Kamp, *Orthogonal polynomial matrices on the unit circle*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 135–160.

[5] R. Freund and T. Huckle, *Extension problem for H-unitary matrices*, Linear Algebra Appl., 108 (1988), pp. 213–230.

[6] B. Fritzsche and B. Kirstein, *An extension problem for nonnegative Hermitian block-Toeplitz matrices*, Math. Nach., 130 (1978), pp. 121–135; Part II, Math. Nach., 131 (1987), pp. 287-297; Part III, Math. Nach., 135 (1988), pp. 319-331; Part IV, Math. Nach., 143 (1989), pp. 329-354; Part V, Math. Nach., 144 (1989), pp. 283-308.

[7] F. Frobenius, *Ableitung einen Satz von C. Caratheodory aus einer Formel von Kronecker*, Sitzungsber. Kgl.-Preuss. Akad. Wiss., (1912), pp. 16–31.

[8] I. Gohberg and H. Dym, *Extensions of matrix valued functions with rational polynomial inverses*, Integral Equations Operator Theory, 2 (1979), pp. 503–528.

[9] I. Gohberg, P. Lancaster and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.

[10] ——, *Matrices and Indefinite Scalar Products*, in Operator Theory: Advances and Applications, Vol. 8, Birkhäuser-Verlag, Basel, 1983.

[11] I. Gohberg and L. Lerer, *Matrix Generalizations of M. G. Krein Theorems on Orthogonal Polynomials*, in Operator Theory: Advances and Applications, OT34, Birkhäuser-Verlag, Basel, 1988.

[12] I. S. Iochvidov, *Hankel and Toeplitz Matrices and Forms*, Birkhäuser-Verlag, Basel, 1982.

[13] P. Lancaster and M. Tismenetsky, *The Theory of Matrices with Applications*, 2nd ed., Academic Press, New York, 1985.

[14] L. Lancaster, L. Lerer, and M. Tismenetsky, *Factored forms of solutions of the equation $AX - XB = C$ in matrices*, Linear Algebra Appl., 62 (1983), pp. 19–49.

[15] I. Schur, *Ueber einen Satz von C. Caratheodory*, Sitzungsber. Kgl.-Preuss. Akad. Wiss., (1912), pp. 4–15.

[16] G. Szego, *Ueber einen Satz von C. Caratheodory*, Jahresber. DMV, 28 (1919), pp. 131–137.

[17] M. Tismenetsky, *Factorizations of hermitian block Hankel matrices*, Linear Algebra Appl., 166 (1992), pp. 45–64.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 113-117, January 1993

009

THE ADDITIVE INVERSE EIGENVALUE PROBLEM FOR LIE PERTURBATIONS*

CHRISTOPHER I. BYRNES† AND XIAOCHANG WANG‡

Abstract. Motivated by several examples arising in linear systems theory, the problem considered here is the inverse eigenvalue problem for an arbitrary square matrix and for arbitrary additive perturbations belonging to a matrix Lie algebra. For an algebraically closed field with characteristic zero, the main theorem gives necessary and sufficient conditions for the positive solution of the corresponding additive inverse eigenvalue problem. There are, of course, several antecedents of this result in the literature involving special coordinate systems and special representations of particular matrix Lie algebras, most notable among these being the result due to S. Friedland on the inverse eigenvalue problem for diagonal perturbations.

Key words. additive eigenvalue problem, Lie algebras, linear control system, algebraically closed fields

AMS(MOS) subject classification. 15A18

1. Introduction.

The problem we consider here is the additive inverse eigenvalue problem over an algebraically closed field F of characteristic zero where the perturbation set has a very useful structure which is present in several motivating examples. We study the problem: Given a matrix $A \in g\ell(n, F)$, can we adjust the eigenvalues of A to an arbitrary set of n complex numbers, counted with multiplicity, by adding a matrix belonging to a given Lie algebra? More precisely, given $A \in g\ell(n, F)$, and a matrix Lie subalgebra $\mathcal{L} \subset g\ell(n, F)$, define a polynomial mapping χ_A from $\mathcal{L}$ to the affine n space $\mathbf{A}^n$ of monic polynomials over F of degree n, via

$$\chi_A(L) = \det(sI - A - L), \qquad L \in \mathcal{L}.$$

In this geometric language, we want to know if χ_A is onto.

Lie perturbations arise in many particular applications. Here, we give some examples arising in control theory and in the matrix analysis of certain inverse spectral problems.

Example 1. Consider a finite-dimensional, time-invariant linear control system

$$\dot{x} = Ax + Bu, \qquad y = Cx$$

in a state space representation which evolves on $\mathbb{R}^m$. Thus the state of the system is denoted by the n-vector $x(t)$ and knowledge of the "initial state" $x(0)$ and the control $u(t)$, where $u(t) \in \mathbb{R}^m$, is sufficient to determine the future state of the system $x(t)$ at any time $t > 0$. The system output $y(t)$ is assumed to be a p-vector, representing either a measured or a controlled variable. One of the principal uses of state feedback, $u = Kx$, is to alter or prescribe the natural frequencies of the control system. Thus an important problem is to determine the possible spectra of the matrices obtained using feedback, where x is an n-vector, u is an m-vector, y is a p-vector.

If we use state feedback $u = Kx$, the closed loop system is

$$\dot{x} = (A + BK)x, \qquad y = Cx.$$

* Received by the editors February 1, 1989; accepted for publication (in revised form) June 6, 1991.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130 (byrnes@cec1.wustl.edu). This research was supported in part by Air Force Office of Scientific Research grant 88-0309 and National Science Foundation grant DMS-9008223.

‡ Department of Mathematics, Texas Tech University, Lubbock, Texas 79409 (mdxia@ttacsl.ttu.edu).

Here the perturbation belongs to $\{BK\}$, B is a fixed $n \times m$ constant matrix, and $\{BK\}$ is a Lie algebra under the commutator product $[X, Y] = XY - YX$ of two matrices. It is a well-known result that χ_A is onto if and only if (A, B) is controllable [3].

Example 2. In some cases of interest, only output measurements are available so that state feedback must be implemented either using dynamically updated estimates $\hat{x}$ of the state x or using the functional Cx of x. For the system of Example 1, if we use output feedback $u = Ky$ the closed loop system is

$$\dot{x} = (A + BKC)x, \qquad y = Cx.$$

Again, $\{BKC\}$ is a Lie algebra, but the problem of arbitrarily "tuning" the natural frequencies of the system by output feedback is still open. For this Lie algebra, however, it is known that χ_A is onto if $mp = n$ and the system is nondegenerate [2].

Example 3. For the same system, if we use the dynamic compensator

$$\dot{z} = Fz + Gy, \qquad u = Hz + Ky,$$

where z is a q-vector, the closed loop system is

$$\begin{pmatrix} x \\ z \end{pmatrix} = \begin{pmatrix} A + BKC & BH \\ GC & F \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix}.$$

Here, for fixed matrices B, C, the subspace of matrices

$$\left\{ \begin{pmatrix} BKC & BH \\ GC & F \end{pmatrix} \right\}$$

is also a Lie algebra.

Example 4. Given a matrix A, if we change the diagonal elements, the perturbation belongs to

$$\mathscr{D}_n = \{\text{diagonal matrices}\},$$

which is a Lie algebra. For this Lie algebra, Friedland has shown that χ_A is onto for all A [1], [5], [6].

In this paper, we are looking for conditions on a Lie algebra ensuring that χ_A is onto for all A. Our main result asserts that χ_A is onto for all A if and only if the rank of the Lie algebra is greater than n, and there exists at least one element in the Lie algebra which has n distinct eigenvalues. The more difficult case of determining whether χ_A is onto for generic A, as in Examples 1-3, can, however, be analyzed using more sophisticated geometric methods (see, e.g., [2], [10], and [11]).

Caveat. Before stating and proving our main result, we want to first clarify what we mean by the rank of a Lie algebra $\mathscr{L}$, since there have been two definitions arising in the literature (compare, e.g., [4] and [8]). A Cartan subalgebra of $\mathscr{L}$ is a Lie subalgebra $h \subset \mathscr{L}$ which is nilpotent and maximal with respect to this property. Since F is algebraically closed, all Cartan subalgebras are conjugate in $\mathscr{L}$ [8] and by the rank of $\mathscr{L}$ we mean $\dim_F h$, as in [4]. An alternative convention, *which we do not follow here*, is to define the rank of $\mathscr{L}$ as the codimension of h in $\mathscr{L}$, as in [8].

As an example, consider the Lie algebra $g\ell(n, F)$ of all $n \times n$ matrices defined over F, which has as a Cartan subalgebra D_n of all diagonal matrices. Thus the rank of $g\ell(n, F)$ is n. Of course, by additively perturbing any matrix A by $g\ell(n, F)$, we can make the perturbed matrix, and hence the perturbed characteristic polynomial, arbitrary. Our main theorem asserts that complete eigenvalue assignment holds precisely because $g\ell(n, F)$ is a matrix algebra having $\mathscr{D}_n$ as a Cartan subalgebra. The result for $\mathscr{D}_n$ itself is a by-now classical result due to Friedland. However, we note that even in this case, our main theorem has the advantage of giving a coordinate-free criterion for a problem that is, of course, stated in a coordinate-free fashion. In § 2, we state Friedland's theorem for $\mathscr{D}_n$ and give a self-contained proof similar to the proofs of surjectivity for the map χ described in Example 2 and developed in algebraic geometric system theory [2]–[11], a proof we feel might be of independent interest.

2. Friedland's theorem. The following theorem is a well-known result due to Friedland.

THEOREM 1 (Friedland). *$\chi_A : \mathscr{D}_n \twoheadrightarrow A^n$ is onto for all $A \in g\ell(n, F)$.*

There are a variety of geometric and topological proofs for this theorem [1], [5], [6]. The proof we give here is based on the projective dimension theorem. Recall that for two varieties Y, Z of dimension r, s in $\mathbf{P}^n$, if $r + s - n \geqq 0$, then $Y \cap Z$ is nonempty and every irreducible component of $Y \cap Z$ has dimension $\geqq r + s - n$ [7].

Proof of Theorem 1. We can write $\chi_A(D)$ as

$$\chi_A(D) = \det \begin{bmatrix} I & I \\ D & sI - A \end{bmatrix}.$$

We define an extension of χ_A as a map $\tilde{\chi}_A = \mathbf{P}_1^1 \times \cdots \times \mathbf{P}_n^1 \twoheadrightarrow \mathbf{P}^n$.

First, denote the ordered diagonal entries of $\mathscr{D}$ by $(y_1, \ldots, y_n)$. We introduce the homogeneous coordinates $[x_i, y_i]$ by rescaling the variables in $\mathscr{D}$ via y_i/x_i and clearing denominators, leading to the function

$$\tilde{\chi}_A([x_1, y_1], [x_2, y_2], \ldots, [x_n, y_n]) = \det \begin{bmatrix} x_1 & & & \\ & \ddots & & I \\ & & x_n & \\ y_1 & & & \\ & \ddots & & sI - A \\ & & y_n & \end{bmatrix},$$

where $\mathbf{P}_i^1$ are one-dimensional projective and $[x_i, y_i]$ are the homogeneous coordinates of $\mathbf{P}_i^1$.

$\tilde{\chi}_A$ is a well-defined morphism and is an extension of χ_A; i.e., $\tilde{\chi}_A$ restricted to the affine space $\mathbf{A}^1 \times \cdots \times \mathbf{A}^1$ with coordinates $[1, y_i]$ coincides with χ_A. More explicitly, the points of $\mathscr{D}_n$ in $\mathbf{P}_1^1 \times \cdots \times \mathbf{P}_n^1$ are just the "finite" points—the points with $x_i \neq 0$, $i = 1, 2, \ldots, n$. χ_A maps the "infinite" points to polynomials of degree less than n, which are exactly the "infinite" points of $\mathbf{P}^n$. Therefore,

$$\chi_A(\mathscr{D}_n) = \tilde{\chi}_A(\mathscr{D}_n) = \tilde{\chi}_A(\mathbf{P}_1^1 \times \cdots \times \mathbf{P}_n^1) \cap \mathbf{A}^n.$$

Since the image of a projective variety under a morphism is closed, if we can prove that the image of $\tilde{\chi}_A$ contains an open set of $\mathbf{P}^n$, then $\tilde{\chi}_A$ must be onto, so that

$$\chi_A(\mathscr{D}_n) = \tilde{\chi}_A(\mathbf{P}_1^1 \times \cdots \times \mathbf{P}_n^1) \cap \mathbf{A}^n = \mathbf{A}^n.$$

For any n distinct elements $s_1, s_2, \ldots, s_n$ in F_1, let $\alpha(x_i)$ be the hypersurface of codimension 1 in $\mathbf{P}_1^1 \times \cdots \times \mathbf{P}_n^1$ defined by

$$\det \begin{bmatrix} x_1 & & & \\ & \ddots & & I \\ & & x_n & \\ y_1 & & & \\ & \ddots & & s_i I - A \\ & & y_n & \end{bmatrix} = 0.$$

Using the projective dimension theorem n times, we deduce that

$$\bigcap_{i=1}^{n} \alpha(s_i) = \tilde{\chi}_A^{-1}[(s - s_1)(s - s_2)\cdots(s - s_n)] \neq \varnothing.$$

Therefore, the open set

$$\{(s - s_1)(s - s_2)\cdots(s - s_n) \mid s_1 \cdots s_n \text{ are distinct}\}$$

is contained in $\tilde{\chi}_A(\mathbf{P}_1^1 \times \cdots \times \mathbf{P}_n^1)$, and the theorem is proved. □

Remark. As we noted in the introduction, the characteristic polynomial of a matrix is coordinate-free, of course, and the set $g\ell(n, F)$ of all matrices A does not depend on a choice of basis, yet the hypothesis of Theorem 1 is coordinate-dependent, in the following sense. A coordinate-free version of Theorem 1 would be the assertion that $\chi_A : \mathscr{L} \to \mathbf{A}^n$ is onto for all A if $\mathscr{L}$ is conjugate to $\mathscr{D}_n$. In addition, a straightforward extension of this result would be the assertion that if for any matrix Lie algebra $\mathscr{L}$ which contains a subalgebra $\mathscr{L}_1$ conjugate to $\mathscr{D}_n$, the map $\chi_A : \mathscr{L} \to \mathbf{A}^n$ is surjective. Our main result shows that this condition is also necessary.

3. Main result.

THEOREM 2. *Given a Lie algebra $\mathscr{L} \in g\ell(n, F)$, $\chi_A : \mathscr{L} \to \mathbf{A}^n$ is onto for all A if and only if rank $\mathscr{L} = n$ and some element of $\mathscr{L}$ has distinct eigenvalues.*

Before we prove Theorem 2, we recall some basic definitions and results concerning Lie algebras [4], [8], [9].

A Lie algebra is called nilpotent if $\mathscr{L}_{(n)} = 0$ for some n, where

$$\mathscr{L}_{(1)} = [\mathscr{L}, \mathscr{L}], \qquad \mathscr{L}_{(k)} = [\mathscr{L}, \mathscr{L}_{(k-1)}].$$

By Engel's theorem, every nilpotent subalgebra of $g\ell(n, F)$ is conjugate to a subalgebra of $s(m_1 \cdots m_r)$ for some choice of $m_1 \cdots m_r$ [8], [9] where $s(m_1 \cdots m_r)$ is the Lie algebra of all block diagonal matrices of

$$\begin{bmatrix} M_1 & & \\ & \ddots & \\ & & M_r \end{bmatrix},$$

where

$$M_i = \begin{pmatrix} \lambda_i & & * \\ & \ddots & \\ 0 & & \lambda_i \end{pmatrix}$$

is an $m_i \times m_i$ upper triangular matrix with the same diagonal elements.

A Cartan subalgebra $\mathscr{A}$ of $\mathscr{L}$ is a nilpotent subalgebra that coincides with its normalizer, i.e.,

$$\mathscr{A} = N(\mathscr{A}) = \{X : X \in \mathscr{L}, \text{ad } X(\mathscr{A}) \subset \mathscr{A}\}.$$

All Cartan subalgebras of $\mathscr{L}$ are conjugate to each other under an automorphism, $L \to GLG^{-1}$, for some G belonging to the corresponding matrix Lie group [4], [8]. The rank of a Lie algebra is the dimension of a Cartan subalgebra.

Finally, we recall that an element $X \in \mathscr{L}$ is called regular if the dimension of $\mathscr{L}_0(X) = \{Y : Y \in \mathscr{L}, (\text{ad } X)^k(Y) = 0 \text{ for some } k\}$ is minimum. In this case, $\mathscr{L}_0(X)$ is a Cartan subalgebra. It is known [8] that regular elements are dense in $\mathscr{L}$.

LEMMA 1. *For a Lie subalgebra $\mathscr{L} \subset gL(n, F)$, rank $\mathscr{L} = n$ and some element of $\mathscr{L}$ has n distinct eigenvalues if and only if any Cartan subalgebra $\mathscr{A} \subseteq \mathscr{L}$ can be written as*

$$\mathscr{A} = G\mathscr{D}_n G^{-1}, \qquad G \in GL(n, F).$$

Proof. We only need to prove the "only if" part. Since the regular elements are dense in $\mathscr{L}$, there exists a regular element X which has n distinct eigenvalues.

Let $\mathscr{A} = \mathscr{L}_0(X) = \{Y : Y \in \mathscr{L}, (\operatorname{ad} X)^k(Y) = 0\}$. By Engel's theorem, $\mathscr{A}$ is a Cartan subalgebra and it is conjugate to a subalgebra of $s(m_1 \cdots m_r)$. To say X has distinct eigenvalues is to say that $r = n$, $m_1 = m_2 = \cdots = m_n = 1$; i.e., $\mathscr{A}$ is conjugate to subalgebra of $\mathscr{D}_n$. But

$$\dim \mathscr{A} = \operatorname{rank} \mathscr{L} = n,$$

so $\mathscr{A}$ is conjugate to $\mathscr{D}_n$. Since any Cartan subalgebra is conjugate to $\mathscr{A}$, any Cartan subalgebra is conjugate to $\mathscr{D}_n$.

LEMMA 2. *Given a Cartan subalgebra $\mathscr{A} \subset \mathscr{L}$, then $U(\mathscr{A}) = \{L \in \mathscr{L} : GLG^{-1} \in \mathscr{A}$ for some $G \in GL(n)\}$ is dense in $\mathscr{L}$.*

Proof. Take any regular element X, then $G\mathscr{L}_0(X)G^{-1} = \mathscr{A}$ for some $G \in GL(n)$. This means that $U(\mathscr{A})$ contains all regular elements, so it is dense in $\mathscr{L}$. □

Proof of Theorem 2. Sufficiency follows from Lemma 1 and Theorem 1. As for necessity, χ_0 is onto, so that

$$\mathbf{A}^n = \chi(\mathscr{L}) = \chi(\overline{U(\mathscr{A})}) \subset \overline{\chi(U(\mathscr{A}))} = \overline{\chi(\mathscr{A})}.$$

But $\mathscr{A}$ is conjugate to a subalgebra of $s(m_1 \cdots m_r)$; $\chi(\mathscr{A}) = \{(s - \lambda_1)^{m_1} \cdots (s - \lambda_r)^{m_r}\}$. $\overline{\chi(\mathscr{A})} = \mathbf{A}^n$ means that $r = n$, $m_1 = \cdots = m_n = 1$, and $\mathscr{A}$ is conjugate to $\mathscr{D}_n$. By Lemma 1, rank $\mathscr{L} = n$, and there exist elements having n distinct eigenvalues.

Examples 1 *and* 2 (bis). As an example, we consider the Cartan subalgebra arising in Examples 1 and 2. The Cartan subalgebra of BK is (conjugate to)

$$\mathscr{D}_m = \{m \times m \text{ diagonal matrices}\},$$

where m is the rank of B. So $\chi : \{BK\} \to \mathbf{A}^n$ is onto for all A if and only if $m = n$.

Similarly, the Cartan subalgebra of BKC is conjugate to the direct sum of $\mathscr{D}_s$ and $M_{(m-s)\times(p-s)}$. Therefore, $\chi : \{BKC\} \to \mathbf{A}^n$ is onto for all A if and only if $s = n$.

REFERENCES

[1] J. C. ALEXANDER, *Matrices, eigenvalues and complex projective space*, Amer. Math. Monthly, 85 (1978), pp. 727–733.

[2] R. W. BROCKETT AND C. I. BYRNES, *Multivariate Nyquist criteria, root loci and pole placement: A geometric viewpoint*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 271–283.

[3] C. T. CHEN, *Linear System Theory and Design*, CBS College Publishing, New York, 1984.

[4] Y. CHEN, *General Theory of Lie Algebras*, Gordon and Breach, New York, 1978.

[5] S. FRIEDLAND, *Matrices with prescribed off-diagonal elements*, Israel J. Math., 11 (1972), pp. 184–189.

[6] ———, *Inverse eigenvalue problems*, Linear Algebra Appl., 17 (1977), pp. 15–51.

[7] R. HARTSHORNE, *Algebraic Geometry*, Springer-Verlag, New York, 1977.

[8] N. JACOBSON, *Lie Algebras*, Dover, New York, 1979.

[9] D. H. SATTINGER AND O. L. WEAVER, *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics*, Springer-Verlag, New York, 1986.

[10] X. WANG, *Geometric inverse eigenvalue problem*, in Computation and Control, K. Bowers and J. Lund, eds., Birkhäuser, Boston, 1988, pp. 375–383.

[11] X. WANG, *Additive Inverse Eigenvalue Problems and Pole Placement of Linear Systems*, Ph.D. thesis, Arizona State University, Tempe, AZ, 1989.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 118-121, January 1993

010

DECREASING THE DISPLACEMENT RANK OF A MATRIX*

VICTOR PAN†

Abstract. This paper approximates a matrix B having a displacement rank r by a matrix B_d having a lower displacement rank d (for a fixed $d < r$), so as to minimize the approximation error norm. $\|B_d - C\|_2$ for any matrix C in terms of $\|B - C\|_2$ is estimated.

Key words. matrix rank, displacement rank, displacement generator

AMS(MOS) subject classifications. 15A57, 65F30, 15A99, 65F99

1. Introduction. The downshift (or displacement) matrix Z (zero everywhere except for its first subdiagonal, filled with ones) shifts down the entries of any vector $\mathbf{v} = [v_0, \dots, v_k]^T$ (so that $Z\mathbf{v} = [0, v_0, \dots, v_{k-1}]$) and plays an important role in the study of Toeplitz-like matrices. In particular, a pair of $n \times d$ matrices G and H is called a *displacement generator* (of length d) for an $n \times n$ matrix S if

$$W(S) = S - ZSZ^T = GH^T, \tag{1.1}$$

and $d(S)$, the minimum d in all such representations of S, is called the displacement rank of the matrix S and characterizes the Toeplitz-like structure of S (see [KVM], [KKM], and [CKL-A]). In particular, $d(S) \leqq 2$ if S or S^{-1} is a Toeplitz matrix, and $d(S) \leqq p + q$ if S or S^{-1} is a $p \times q$ block matrix with Toeplitz blocks.

We may represent S by G and H (see (3.1) below), thus saving $(n - 2d)n$ words of memory for a smaller d and accelerating the subsequent computations with S. Moreover, given a matrix, we may wish to work with a neighboring matrix having a shorter displacement generator (compare Remark 3.2 below). This leads us to the following problem.

Problem 1 (decreasing the length of a displacement generator of a matrix). Let a pair of given $n \times r$ matrices M and N form a displacement generator (M, N) of length r for an $n \times n$ matrix B. Then

(a) find a displacement generator (G_d, H_d) of length d for some matrix B_d such that the norm $\|MN^T - G_dH_d^T\|_2$ is minimum where d is a fixed integer, $0 \leqq d \leqq r$;

(b) for an $n \times n$ matrix C having displacement rank $d < r$, estimate the norm $\|B_d - C\|_2$ as a function in $\Delta = \|B - C\|_2$.

Thus, in part (a) we seek approximation to B by B_d with $d(B_d) \leqq d$ that minimizes $\|W(B) - W(B_d)\|_2$ ($W(S)$ being defined by (1.1)).

This does not generally give us a matrix C_d that minimizes $\|C_d - B\|_2$ subject to $d(C_d) \leqq d$, and we do not know if such a matrix C_d is readily available, but our solution to part (b) (for $C = C_d$) shows that $\|B_d - B\|_2 \leqq \|B_d - C_d\|_2 + \|B - C_d\|_2$ is never greater than $(2 + 2n(r - d))\|B - C_d\|_2$.

We solve Problem 1 in § 3 by using the known auxiliary techniques that we recall in the next section.

* Received by the editors June 12, 1990; accepted for publication (in revised form) May 15, 1991. This research was supported by National Science Foundation grants CCR-8805782 and CCR-9020690 and by Professional Staff Congress of City University of New York awards 661340, 668541, and 669290.

† Department of Mathematics and Computer Science, Lehman College, The City University of New York, Bronx, New York 10468 and Computer Science Department, State University of New York at Albany, Albany, New York 12222 (pan@cs.albany.edu).

Our solution can be immediately extended to the case of other representations of structured matrices based on the application of operators of scaling, displacement (shift), and other transformations (cf. [GKK] and [P89a]).

2. Approximation by a lower rank matrix. An $n \times n$ matrix W of rank $r \leqq n$ has many distinct decompositions of the form

$$W = MN^T, \quad M \in \mathbf{R}_{n,r}, \quad N \in \mathbf{R}_{n,r}. \tag{2.1}$$

The singular value decomposition (SVD) of W,

$$W = U\Sigma^2 V^T, \quad U \in \mathbf{R}_{n,r}, \quad V \in \mathbf{R}_{n,r}, \quad U^TU = V^TV = I_r, \tag{2.2}$$

$$\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_r), \qquad \sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_r > 0 \tag{2.3}$$

represents a special case of the decomposition (2.1) where $M = G$, $N = H$,

$$W = GH^T, \quad G = U\Sigma, \quad H = V\Sigma. \tag{2.4}$$

The matrix W uniquely determines the matrix Σ of the relations (2.2)–(2.3) and, if all the singular values σ_k^2 are distinct, it uniquely defines the matrices U and V as well. For an integer d such that $0 \leqq d \leqq r$, let U_d, V_d, G_d, and H_d denote the four matrices formed by the first d columns of the four matrices U, V, G, and H, respectively. We write

$$W_d = G_dH_d^T, \quad G_d = U_d\Sigma_d, \quad H_d = V_d\Sigma_d, \quad \Sigma_d = \operatorname{diag}(\sigma_1, \dots, \sigma_d). \tag{2.5}$$

Then W_d has rank d, and (see [GL, p. 73])

$$\|W - W_d\|_2 \leqq \|W - Y\|_2 \quad \text{for all matrices } Y \text{ of rank at most } d. \tag{2.6}$$

Now suppose that we are given two matrices M and N of the decomposition (2.1) and an integer d such that $0 \leqq d \leqq r$, and that we seek the matrices G_d and H_d satisfying (2.1)–(2.6). Surely, we may solve this problem by means of computing the SVD (2.2), but we may also decrease the size of the problem and therefore the overall computational cost by first computing (recursive) QR-factorization of the $n \times r$ matrices $M = Q(M)R(M)$ and $N = Q(N)R^T(N)$ (see [GL, pp. 211–213]). The matrices $R(M)$ and $R(N)$ overwrite M and N, respectively, and the factors $Q(M)$ and $Q(N)$ are not accumulated, but are stored as the products of Householder matrices. The computation is reduced to a sequence of Householder transformations and is performed by using about $2r^2(n - \frac{r}{3})$ flops. Then it remains to compute the SVD of the $r \times r$ matrix $R(M)R(N)$; actually, we only need the d largest singular values $\sigma_1^2, \dots, \sigma_d^2$ and the first d columns of the matrices G and H. We may apply various effective iterative algorithms, which are numerically stable and rapidly convergent. This stage is relatively inexpensive if r is much less than n, which is usually the case where we deal with Toeplitz or Toeplitz-like matrices. For instance, about $21r^3$ flops usually suffice to yield the solution within the machine precision if we apply the old approach; that is, if we first compute the product $R(M)R(N)$ and then its SVD by using algorithms essentially reduced to a sequence of Givens rotations (see [GL, p. 239]). We refer the reader to [EL], [FH], and [HLPW] for the more recent effective and numerically stable approach (based on the Jacobi method) to computing the SVD of the product $R(M)R(N)$ without explicitly computing the product itself.

The overall computation requires about $2nr^2 + O(r^3)$ flops and allows a substantial parallel acceleration. (From a theoretical point of view, this computation can be placed in NC, even where $r = n$; see [P88b].)

Remark 2.1. Clearly, if W is symmetric, then $U = V$ and W_d is symmetric too, but the transition to W_d does not generally preserve the persymmetricity of W.

3. Decreasing the length of a displacement generator. To solve Problem 1(a), we apply the algorithm of the previous section to the matrix MN^T. We refer to (2.6) for substantiation and to the previous section for the computational cost estimates.

The meaning of this solution is better understood when we also solve Problem 1(b), by relying on the *displacement representation*

$$S = S(G, H) = \sum_{i=1}^{d} L(\mathbf{g}_i)L^T(\mathbf{h}_i), \tag{3.1}$$

which is equivalent to (1.1), provided that $G = [\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_d]$, $H = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_d]$, and that $L(\mathbf{v})$ denotes the triangular Toeplitz matrix defined by its first column $\mathbf{v}$. (This representation, due to [KKM], is actually fundamental to the study of Toeplitz-like matrices.)

To estimate the norm $\| B_d - C\|_2$, we apply (3.1) to the matrix $S = B$, replace d by r in (3.1), and obtain that

$$B = B(G, H) = \sum_{i=1}^{r} L(\mathbf{g}_i)L^T(\mathbf{h}_i), \tag{3.2}$$

$$B_d = B(G_d, H_d) = \sum_{i=1}^{d} L(\mathbf{g}_i)L^T(\mathbf{h}_i), \tag{3.3}$$

where $\mathbf{g}_i$ and $\mathbf{h}_i$ denote columns i of the matrices G and H, respectively, for $i = 1, \dots, r$, which are also columns i of the matrices G_d and H_d, respectively, for $i = 1, \dots, d$ (compare (2.4) and (2.5)). Equations (2.2)-(2.4) imply that

$$\|\mathbf{g}_i\|_2 = \|\mathbf{h}_i\|_2 = \sigma_i \quad \text{for } i = 1, \dots, r. \tag{3.4}$$

It follows from (3.2) and (3.3) that

$$\|B - B_d\|_2 = \left\| \sum_{i=d+1}^{r} L(\mathbf{g}_i)L^T(\mathbf{h}_i) \right\|_2 \leqq \sum_{i=d+1}^{r} \|L(\mathbf{g}_i)L^T(\mathbf{h}_i)\|_2 .$$

Apply (3.4) and obtain

$$\|B - B_d\|_2 \leqq n \sum_{i=d+1}^{r} \sigma_i^2 . \tag{3.5}$$

We are prepared to deduce the following result.

PROPOSITION 3.1. *Let the matrices B, B_d and C, and a positive scalar Δ be defined as in Problem* 1. *Then* $\|C - B_d\|_2 \leqq (1 + 2n(r - d))\Delta$.

Proof. Apply (1.1) for $S = B$ and $S = C$ to define the matrices $W(B)$ and $W(C)$ and deduce that

$$\|W(B) - W(C)\|_2 \leqq 2\Delta. \tag{3.6}$$

Let $\sigma_i^2(B)$ and $\sigma_i^2(C)$ denote the ith singular values of the matrices $W(B)$ and $W(C)$, respectively, so that $\sigma_1(B) \geqq \sigma_2(B) \geqq \cdots \geqq \sigma_d(B) \geqq 0$, $\sigma_1(C) \geqq \cdots \geqq \sigma_d(C) \geqq 0$ and

$$\sigma_i(C) = 0 \quad \text{if } i > d. \tag{3.7}$$

Apply the well-known estimate [GL, p. 428] and deduce that

$$|\sigma_i^2(B) - \sigma_i^2(C)| \leqq \|W(B) - W(C)\|_2 \quad \forall i.$$

Substitute the relations (3.6) and (3.7) and obtain $\sigma_i^2(B) \leqq 2\Delta$ if $i > d$, so that $\sum_{i=d+1}^{r} \sigma_i^2(B) \leqq 2(r-d)\Delta$. Substitute this bound into the inequality (3.5) and deduce that

$$\|C - B_d\|_2 \leqq \|C - B\|_2 + \|B - B_d\|_2 \leqq (1 + 2n(r-d))\Delta. \qquad \square$$

Remark 3.1. $B_d = B$ if $\Delta = 0$, so that we have an algorithm that computes a displacement generator of length $d(A)$ for a matrix A given with its (longer) displacement generators of $d > d(A)$. This problem has other effective solutions (see, for instance, [CKL-A] and [P90a]). Our present solution has the advantage of being both simple for smaller d and numerically stable.

Remark 3.2. Originally, the author encountered Problem 1 while trying to approximate to the inverse A^{-1} of a Toeplitz-like matrix A. The displacement ranks of the computed approximations to A^{-1} grew (even though $d(A^{-1})$ was small), which complicated the computations and thus lead to Problem 1.

Acknowledgments. The author wishes to thank Dr. Thomas Kailath for the discussion that inspired this paper, and the referees and the editor Dr. G. Cybenko for helpful comments. He also thanks Joan Bentley for her excellent assistance with typing the manuscript.

REFERENCES

[CKL-A] J. Chun, T. Kailath, and H. Lev-Ari, *Fast parallel algorithm for* QR-*factorization of structured matrices*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 899–913.

[EL] M. Ewerbring and F. T. Luk, *Canonical correlations and generalized* SVD: *Applications and new algorithms*, J. Comput. Appl. Math., 27 (1989), pp. 37–52.

[FH] K. V. Fernando and S. J. Hammarling, *A product induced singular value decomposition for two matrices and balanced realisation*, in Linear Algebra in Signals, Systems and Control, B. N. Datta, C. R. Johnson, M. A. Kaashoek, R. Plemmons, and E. Sontag, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1988, pp. 128–140.

[GKK] I. Gohberg, T. Kailath, and I. Koltracht, *Efficient solution of linear systems of equations with recursive structure*, Linear Algebra Appl., 80 (1986), pp. 81–113.

[GL] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1989.

[HLPW] M. T. Heath, A. J. Laub, C. C. Paige, and R. C. Ward, *Computing the* SVD *of a product of two matrices*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1147–1159.

[KKM] T. Kailath, S.-Y. Kung, and M. Morf, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[KVM] T. Kailath, A. Viera, and M. Morf, *Inverses of Toeplitz operators, innovations, and orthogonal polynomials*, SIAM Rev., 20 (1978), pp. 106–119.

[P88b] V. Pan, *New methods for computations with Toeplitz-like matrices*, Tech. Rep. 88-28, Computer Science Dept., State Univ. of New York, Albany, NY, 1988.

[P89a] ———, *On some computations with dense structured matrices*, in Proc. ACM–SIGSAM Internat. Sympos. on Symbolic and Alg. Comp., 1989, pp. 34–42; Math. Comp., 55 (1990), pp. 179–190.

[P90a] ———, *Parameterization of Newton's iteration for computations with structured matrices and applications*, Tech. Rep. CUCS-032-90, Computer Science Dept., Columbia Univ., New York; Comp. Math. Appl., 3 (1992), pp. 61–75.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 122-131, January 1993

011

ON THE CONVERGENCE OF THE MSOR METHOD FOR SOME CLASSES OF MATRICES*

D. HERCEG†, M. M. MARTINS‡, AND M. E. TRIGO§

Abstract. This paper is concerned with the solution of the linear system $Ax = b$ by the modified successive overrelaxation (MSOR) method, if A belongs to some classes of matrices. It is proven that the classes presented here are subclasses of H-matrices. Since, the general conditions for convergence of MSOR method for the class of H-matrices are not always easy to check in practice, some more practical conditions concerning these subclasses of H-matrices are given in this paper.

Key words. linear systems, MSOR method

AMS(MOS) subject classification. 65F10

1. Introduction. Let us consider the linear system

$$\tilde{A}x = \tilde{b},$$

where $x, \tilde{b} \in \mathbf{C}^n$ with $\tilde{b}$ known and x unknown, $\tilde{A} \in \mathbf{C}^{n,n}$ has the property $\mathscr{A}$, i.e., $\tilde{A}$ is diagonal or there exists a permutation matrix P such that $A = P^{-1}\tilde{A}P$ has the form

$$A = \begin{bmatrix} D_1 & T \\ S & D_2 \end{bmatrix}, \tag{1.1}$$

with D_1 and D_2 square diagonal matrices of order k and $n - k$, respectively.

From now on, we consider the linear system

$$Ax = b, \tag{1.2}$$

where A has the form (1.1) and D_1 and D_2 are nonsingular matrices.

To approximate the solution of (1.2), we can use the SOR method with the parameters ω, ω' and from [16] we obtain the modified SOR method (MSOR) given by

$$x^{(i+1)} = M_{\omega,\omega'}x^{(i)} + k_{\omega,\omega'}, \qquad i = 0, 1, 2, \ldots,$$

where

$$M_{\omega,\omega'} = (I - \omega'Q)^{-1}[(1 - \omega)I + \omega B + (\omega - \omega')R]$$

and

$$k_{\omega,\omega'} = -\omega(I - \omega'Q)d,$$

with

$$Q = \begin{bmatrix} 0 & 0 \\ -D_2^{-1}S & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & -D_1^{-1}T \\ 0 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} 0 & 0 \\ 0 & I_2 \end{bmatrix}, \tag{1.3}$$

* Received by the editors November 6, 1989; accepted for publication (in revised form) March 26, 1991. This work was supported by "Instituto Nacional de Investigação Científica."

† Institute of Mathematics, University of Novi Sad, dr.Ilije Djuričića 4, 21000 Novi Sad, Yugoslavia. This work was written while the author was visiting the Department of Mathematics, University of Coimbra, 3000 Coimbra, Portugal.

‡ Department of Mathematics, University of Coimbra, 3000 Coimbra, Portugal.

§ High Current Account and Administration Institute of Coimbra, 3000 Coimbra, Portugal.

and

$$d = \begin{bmatrix} -D_1^{-1} b_1 \\ -\dfrac{\omega'}{\omega} D_2^{-1} b_2 \end{bmatrix},$$

where the vectors b_1 and b_2 are obtained splitting the vector according to the partition used for A.

In [16], Young shows that some variants of the MSOR method are faster than the SOR method. Some results about the convergence of the MSOR method were also presented in [16].

Sufficient convergence conditions for the MSOR method were also presented in [10] and [13], when A of (1.1) is a strictly diagonally dominant matrix and an H-matrix.

Below, we will introduce some subclasses of H-matrices, which were considered in the study of convergence of the linear and nonlinear accelerated overrelaxation (AOR) method [2]-[9].

In [10], [13], and [14] general conditions of convergence for the class of H-matrices were presented, but, in practice, they are not always easy to check. In [10] the class C_1 was also considered.

Therefore, we obtain more practical sufficient convergence conditions for the MSOR method, where the matrix A of (1.1) belongs to some subclasses of H-matrices.

2. Preliminaries and notation. Let $M = [m_{ij}]$ be a real $n \times n$ matrix. We need the following notation

$$\begin{aligned}
N &= \{1, 2, \ldots, n\}, \\
N(i) &= N \setminus \{i\}, \qquad i \in N, \\
N_1 &= \{1, 2, \ldots, k\}, \qquad k \in N, \\
N_2 &= \{k+1, \ldots, n\}, \\
N_1(i) &= N_1 \setminus \{i\}, \qquad i \in N_1, \\
N_2(i) &= N_2 \setminus \{i\}, \qquad i \in N_2, \\
P_i(M) &= \sum_{j \in N(i)} |m_{ij}|, \qquad i \in N, \\
P_i^*(M) &= \sum_{j \in N(i)} |m_{ji}|, \qquad i \in N, \\
P_i'(M) &= \sum_{j=1}^{i-1} |m_{ij}|, \quad i \in N \setminus \{1\}, \quad P_1'(M) = 0, \\
b_i &= P_i(B), \quad q_i = P_i(Q), \quad r_i = P_i(R),
\end{aligned}$$

$$b = \max_{i \in N_1} b_i \quad \text{and} \quad q = \max_{i \in N_2} q_i, \quad \text{with } B,\ Q, \text{ and } R \text{ defined in (1.3).}$$

We use the following definitions and theorems in this paper.

Definition 2.1 [1]. A matrix $A \in \mathbf{C}^{n,n}$ belongs to the class C_0 if and only if A has the form (1.1) and is a strictly diagonally dominant matrix.

Definition 2.2 [12]. A matrix $A \in \mathbf{C}^{n,n}$ belongs to the class C_1 if and only if A has the form (1.1) and there exists $\alpha \in [0, 1]$ such that

$$|a_{ii}| > \alpha P_i(A) + (1 - \alpha) P_i^*(A), \qquad i \in N. \tag{2.1}$$

DEFINITION 2.3 [2]. A matrix $A \in \mathbf{C}^{n,n}$ belongs to the class C_2 if and only if A has the form (1.1) and it holds that

$$|a_{ii}|\,|a_{jj}| > P_i(A)P_j(A), \quad i \in N_1, \quad j \in N_2. \tag{2.2}$$

DEFINITION 2.4 [2]. A matrix $A \in \mathbf{C}^{n,n}$ belongs to the class C_3 if and only if A has the form (1.1) and if there exists $i \in N$ such that

$$|a_{ii}|(|a_{jj}| - P_j(A) + |a_{ji}|) > P_i(A)|a_{ji}|, \qquad j \in N(i). \tag{2.3}$$

DEFINITION 2.5 [1]. A matrix $A \in \mathbf{C}^{n,n}$ is a lower semistrictly diagonally dominant matrix if and only if $|a_{ii}| \geqq P_i(A)$, $i \in N$, and $|a_{ii}| > P'_i(A)$, $i \in N$.

DEFINITION 2.6 [15]. A matrix $A \in \mathbf{R}^{n,n}$ is an M-matrix, if it is nonsingular, $A^{-1} \geqq 0$ and $a_{ij} \leqq 0$, for all $i, j \in N$, $i \neq j$.

DEFINITION 2.7 [16]. A matrix A is an H-matrix if the comparison matrix $M(A)$ defined by $m_{ii} = |a_{ii}|$, $m_{ij} = -|a_{ij}|$, $i, j \in N$, $i \neq j$ is an M-matrix.

THEOREM 2.1 (see [10]). *If* $|\omega'|q_i < 1$, $i \in N_2$, *then the spectral radius of* $L_{\omega,\omega'}$, *i.e.*, $\rho(L_{\omega,\omega'})$, *satisfies*

$$\rho(L_{\omega,\omega'}) \leqq \max\{r_1, r_2\},$$

where

$$r_1 = \max\{|1-\omega| + |\omega| b_i, i \in N_1\},$$

and

$$r_2 = \max\left\{\frac{|1-\omega'|}{1-|\omega'|q_i}, i \in N_2\right\}.$$

THEOREM 2.2 (see [10]). *If* A *belongs to the class* C_0, *then the* MSOR *method converges if*

$$\omega \in \left(0, \frac{2}{1+b}\right) \quad \text{and} \quad \omega' \in \left(0, \frac{2}{1+q}\right).$$

3. Convergence conditions for the MSOR method. First, we prove that the class C_2 is a subclass of H-matrices in the following.

THEOREM 3.1. *If* A *belongs to the class* C_2, *then* A *is an H-matrix.*

Proof. Let $W = \operatorname{diag}(\gamma_1, \ldots, \gamma_k, 1, \ldots, 1)$, with $\gamma_1 = \cdots = \gamma_k = \gamma$, a real parameter, such that $b < \gamma < \frac{1}{q}$.

As A belongs to the class C_2, it is obvious that AW has the form (1.1). From (2.2), we have $b < \frac{1}{q}$ and then we can choose some $\gamma \in (b, \frac{1}{q})$.

Since

$$|(AW)_{ii}| = \begin{cases} \gamma|a_{ii}|, & i \in N_1, \\ |a_{ii}|, & i \in N_2, \end{cases}$$

$$P_i(AW) = \begin{cases} P_i(A), & i \in N_1, \\ \gamma P_i(A), & i \in N_2, \end{cases}$$

and

$$\max\left\{\frac{P_i(A)}{|a_{ii}|}, i \in N_1\right\} < \gamma, \qquad \frac{1}{\gamma} > \max\left\{\frac{P_i(A)}{|a_{ii}|}, i \in N_2\right\},$$

we obtain

$$|(AW)_{ii}| > P_i(AW), \qquad i \in N,$$

i.e., AW belongs to the class C_0.

Then A is a generalized diagonally dominant matrix by rows [11], and by [15] we can say that A is an H-matrix. □

We also need the following theorem.

THEOREM 3.2 (see [13]). *Let $L_{\omega,\omega'}$ and $\hat{L}_{\omega,\omega'}$ be the iteration matrices for the* MSOR *method for A and AW. Then $\hat{L}_{\omega,\omega'}$ and $L_{\omega,\omega'}$ have the same eigenvalues.*

THEOREM 3.3. *Let A be a matrix from the class C_2 and $b < \gamma < \frac{1}{q}$. Then the* MSOR *method is convergent if*

$$(3.1) \qquad \omega \in \left(0, \frac{2\gamma}{b+\gamma}\right) \quad \text{and} \quad \omega' \in \left(0, \frac{2}{1+q\gamma}\right), \quad \text{with} \quad \gamma \in \mathbb{R}.$$

Proof. Let $\hat{L}_{\omega,\omega'}$ be an iteration matrix associated to $W^{-1}AW$, i.e.,

$$\hat{L}_{\omega,\omega'} = (I - \omega'\hat{Q})^{-1}[(1-\omega)I + \omega\hat{B} + (\omega - \omega')\hat{R}] = W^{-1}L_{\omega,\omega'}W,$$

where $\hat{Q} = W^{-1}QW$, $\hat{B} = W^{-1}BW$, and W is the diagonal matrix, defined as in the proof of Theorem 3.1.

It is clear that the matrices AW and $W^{-1}AW$ are both strictly diagonally dominant matrices.

Since $\rho(L_{\omega,\omega'}) = \rho(\hat{L}_{\omega,\omega'})$, from Theorem 2.2 we can say that $\rho(\hat{L}_{\omega,\omega'}) < 1$ if

$$(3.2) \qquad \omega \in \left(0, \frac{2}{1+\hat{b}}\right) \quad \text{and} \quad \omega' \in \left(0, \frac{2}{1+\hat{q}}\right),$$

where $\hat{b} = \max\{\hat{b}_i, i \in N_1\}$ and $\hat{q} = \max\{\hat{q}_i, i \in N_2\}$, with $\hat{b}_i = P_i(\hat{B})$ and $\hat{q}_i = P_i(\hat{Q})$.

It is easy to see that $\hat{b} = \frac{b}{\gamma}$, $\hat{q} = \gamma q$, and from (3.2), we obtain (3.1). Then we get the required result. □

THEOREM 3.4. *Let A be a matrix from class C_2. Then the* MSOR *method converges if*

$$(3.3) \qquad \omega \in \left(0, \frac{2}{1+bq}\right) \quad \textit{and} \quad \omega' \in (0, \phi(\omega)),$$

where

$$\phi(\omega) = \begin{cases} \dfrac{2}{1+bq} & \textit{if } \omega \leqq 1 \\[2ex] \dfrac{2(2-\omega)}{2-\omega(1-bq)} & \textit{if } \omega \geqq 1. \end{cases}$$

Proof. The function $\phi(\omega)$ defined above is a nonincreasing function for $\omega \in (0, (2/1+bq)]$, so we have

$$\phi(\omega) \geqq \phi\left(\frac{2}{1+bq}\right) = 1.$$

Therefore, for ω and ω', which satisfy (3.3), we may consider the following intervals:

$$\Gamma = \begin{cases} \left(b, \dfrac{1}{q}\right) & \text{if } \omega, \omega' \in (0, 1], \\ \left(b, \dfrac{2-\omega'}{q\omega'}\right) & \text{if } \omega \in (0, 1], \quad \omega' \in (1, \phi(\omega)), \\ \left(\dfrac{b\omega}{2-\omega}, \dfrac{1}{q}\right) & \text{if } \omega \in \left(1, \dfrac{2}{1+bq}\right), \quad \omega' \in (0, 1], \\ \left(\dfrac{b\omega}{2-\omega}, \dfrac{2-\omega'}{q\omega'}\right) & \text{if } \omega \in \left(1, \dfrac{2}{1+bq}\right), \quad \omega' \in (1, \phi(\omega)). \end{cases}$$

We prove that the interval Γ is well defined in each of our four cases, that $\Gamma \subset (b, \frac{1}{q})$, and that for any $\gamma \in \Gamma$, the condition (3.1) is satisfied. Then if we use the previous theorem, we can conclude that the MSOR method is convergent.

In the first case, it is obvious that (3.1) is valid for all $\gamma \in (b, \frac{1}{q})$. In the second case we have

$$\phi(\omega) = \frac{2}{1+bq} \quad \text{and} \quad b < \frac{2-\omega'}{q\omega'} < \frac{1}{q},$$

i.e., $\Gamma \subset (b, \frac{1}{q})$. Therefore, for $\gamma \in \Gamma$, we have

$$0 < \omega \leqq 1 < \frac{2\gamma}{b+\gamma} \quad \text{and} \quad 0 < \omega' < \frac{2}{1+q\gamma}.$$

In the third case it holds that $\Gamma \subset (b, \frac{1}{q})$ because

$$b < \frac{b\omega}{2-\omega} < \frac{1}{q}$$

and for $\gamma \in \Gamma$, (3.1) is valid. In the fourth case from $\omega' < \phi(\omega)$, we obtain

$$b < \frac{b\omega}{2-\omega} < \frac{2-\omega'}{q\omega'} < \frac{1}{q}.$$

Then for $\gamma \in \Gamma$ it follows that (3.1) is satisfied. □

COROLLARY 3.1. *If A is a matrix of the form* (1.1) *and lower semistrictly diagonally dominant, then the* MSOR *method converges for ω and ω' given in Theorem* 3.4.

Proof. It is easy to verify that in this case it holds that

$$b_i \leqq 1, \quad i \in N_1, \quad q_j < 1, \quad j \in N_2,$$

and thus $bq < 1$. □

COROLLARY 3.2. *Let A be a matrix of the form* (1.1) *with $S = T^T$, and let*

$$|a_{ii}|\,|a_{jj}| > P_i^*(T)P_j(T), \quad i \in N_1, \quad j \in N_2;$$

then the MSOR *method is convergent for ω and ω' given in Theorem* 3.4, *but with*

$$q = \max\left\{\frac{P_j^*(T)}{|a_{jj}|} : j \in N_2\right\}.$$

COROLLARY 3.3. *Let A be a matrix of the form* (1.1) *with $S = 0$ or $T = 0$, then the* MSOR *method is convergent for $\omega, \omega' \in (0, 2)$.*

We would like to point out that the class C_2 presented here is not the same as that given by

$$|a_{ii}|\,|a_{jj}| > P_i(A)P_j(A), \quad i \in N, \quad j \in N(i),$$

which was considered in [2].

THEOREM 3.5. *If A belongs to the class C_3, then it is an H-matrix.*

Proof. We must find a diagonal matrix W, such that AW is a strictly diagonally dominant matrix. Let $W = \operatorname{diag}(\omega_1, \omega_2, \ldots, \omega_n)$ with $\omega_j = 1, j \in N(i)$, and $\omega_i = \gamma \in \Gamma$, where

$$(3.4) \qquad \Gamma = \begin{cases} \left(\dfrac{P_i(A)}{|a_{ii}|}, \infty\right), & \text{if } a_{ji} = 0, \quad j \in N(i), \\ \left(\dfrac{P_i(A)}{|a_{ii}|}, \gamma_0\right), & \text{if } a_{ji} \neq 0 \quad \text{for some } j \in N(i), \end{cases}$$

$$\gamma_0 = 1 + \min_j \left[\frac{|a_{jj}| - P_j(A)}{|a_{ji}|} : j \in N(i), \qquad a_{ji} \neq 0\right].$$

Since

$$|(AW)_{jj}| = \begin{cases} |a_{jj}|, & j \in N(i), \\ \gamma |a_{ii}|, & j = i, \end{cases}$$

$$P_j(AW) = \begin{cases} P_j(A) + |a_{ji}|(\gamma - 1), & j \in N(i), \\ P_i(A), & j = i, \end{cases}$$

as A belongs to the class C_3, we have

$$|(AW)_{jj}| > P_j(AW), \qquad j \in N,$$

i.e., AW is a generalized diagonally dominant matrix by rows [11]. Now, by [15], we can conclude that A is an H-matrix. □

THEOREM 3.6. *Let A be a matrix from the class C_3 with $i \in N_1$ and $\gamma \in \Gamma$, with Γ from* (3.4). *Then the* MSOR *method is convergent if*

$$(3.5) \qquad \omega \in \left(0, \frac{2}{1+\hat{b}}\right), \qquad \omega' \in \left(0, \frac{2}{1+\hat{q}}\right),$$

where

$$\hat{b} = \max\left[\frac{b_i}{\gamma}, \max\{b_j : j \in N_1(i)\}\right],$$

$$\hat{q} = \max\left[q_j + |q_{ji}|(\gamma - 1) : j \in N_2\right].$$

Proof. Let W be the matrix used in the proof of the previous theorem and $\hat{Q} = W^{-1}QW$, $\hat{B} = W^{-1}BW$.

Since A is a matrix from the class C_3 from (2.3), we have for $i \in N_1$ that

$$1 - b_j > 0, \qquad j \in N_1(i),$$

$$1 - q_j + |q_{ji}|(1 - b_i) > 0, \qquad j \in N_2,$$

and then

$$\gamma_0 = 1 + \min_j \left\{ \frac{1 - q_j}{|q_{ji}|}, j \in N_2, q_{ji} \neq 0 \right\}. \tag{3.6}$$

Now,

$$\hat{b} = \max\{\hat{b}_i : i \in N_1\},$$
$$\hat{q} = \max\{\hat{q}_i : i \in N_2\},$$

and following the proof of Theorem 3.3, we get the required result. □

In the same way we can prove the following theorem.

THEOREM 3.7. *Let A be a matrix from the class C_3 with $i \in N_2$ and let $\gamma \in \Gamma$ where*

$$\Gamma = \begin{cases} (q_i, \infty), & \text{if } b_{ji} = 0, \quad j \in N_2(i), \\ (q_i, \gamma_0), & \text{if } b_{ji} \neq 0 \quad \text{for some } j \in N, \end{cases}$$

$$\gamma_0 = 1 + \min \left\{ \frac{1 - b_j}{|b_{ji}|} : j \in N_1, b_{ji} \neq 0 \right\}.$$

The MSOR *method is convergent if*

$$\omega \in \left(0, \frac{2}{1 + \bar{b}}\right), \qquad \omega' \in \left(0, \frac{2}{1 + \bar{q}}\right),$$

where

$$\bar{b} = \max\{b_j + |b_{ji}|(\gamma - 1) : j \in N_1\},$$

$$\bar{q} = \max\left\{\frac{q_i}{\gamma}, \max\{q_j : j \in N_2(i)\}\right\}.$$

THEOREM 3.8. *Let A be a matrix from the class C_3 with $i \in N_1$ and γ_0 defined by* (3.6). *Then the* MSOR *method converges if*

$$\omega \in \left(0, \frac{1}{1 + \tilde{b}}\right), \qquad \omega' \in (0, \phi(\omega)), \tag{3.7}$$

where

$$\phi(\omega) = \begin{cases} \dfrac{2}{1 + \tilde{q}}, & \text{if } \omega \in (0, 1], \\ \min\left\{ \dfrac{2(2 - \omega)}{(2 - \omega)(1 + q_j - |q_{ji}|) + \omega b_i |q_{ji}|} : j \in N_2 \right\}, & \text{if } \omega \geqq 1, \end{cases}$$

$$\tilde{b} = \max\left\{\frac{b_i}{\gamma_0}, b'\right\}, \qquad b' = \max\{b_j : j \in N_1(i)\},$$

$$\tilde{q} = \max\{q_j + |q_{ji}|(b_i - 1) : j \in N_2\}.$$

Proof. We use the same technique as in the proof of Theorem 3.4. Since A belongs to the class C_3, it holds that $\tilde{b} < 1$, $\tilde{q} < 1$ and $\phi(\omega) > 1$, if $\omega \geqq 1$.

So, for ω and ω', which satisfy (3.7), we may consider intervals

$$G = \begin{cases} \Gamma, & \text{if } \omega, \omega' \in (0, 1], \\ (b_i, \gamma_1(\omega')), & \text{if } \omega \in (0, 1], \quad \omega' \in \left(1, \dfrac{2}{1+\tilde{q}}\right), \\ \left(\dfrac{b_i\omega}{2-\omega}, \infty\right) \cap \Gamma, & \text{if } \omega \in \left(1, \dfrac{2}{1+\tilde{b}}\right), \quad \omega' \in (0, 1], \\ \left(\dfrac{b_i\omega}{2-\omega}, \gamma_1(\omega')\right), & \text{if } \omega \in \left(1, \dfrac{2}{1+\tilde{b}}\right), \quad \omega' \in (1, \phi(\omega)), \end{cases}$$

where Γ is defined by (3.4). Thus

$$\Gamma = \begin{cases} (b_i, \infty), & \text{if } q_{ji} = 0, \quad j \in N_2, \\ (b_i, \gamma_0), & \text{if } q_{ji} \neq 0 \quad \text{for some } j \in N_2, \end{cases}$$

and

$$\gamma_1(\omega') = \begin{cases} \infty \quad \text{if } q_{ji} = 0, \quad j \in N_2, \\ 1 + \min\left[\dfrac{2 - \omega'(1 + q_j)}{\omega'|q_{ji}|} : j \in N_2, \, q_{ji} \neq 0\right]. \end{cases}$$

In the following, we also use $\hat{b}$ and $\hat{q}$, defined in Theorem 3.6. Note that for $\gamma \in \Gamma$ it holds that $\hat{b} < 1$ and $\hat{q} < 1$.

First we prove that interval G is well defined in each of our cases and that $G \subset \Gamma$. Then, for any $\gamma \in G$, we can prove that the condition (3.5) is satisfied and apply Theorem 3.6 to conclude the proof.

In the first case, it is obvious that (3.5) is satisfied for all $\gamma \in \Gamma$. In the second case, if $G = \Gamma$ then (3.5) is also valid for all $\gamma \in \Gamma$, and if $G = (b_i, \gamma_1)$, we have

$$b_i = \gamma_1\left(\frac{2}{1+\tilde{q}}\right) < \gamma_1(\omega') < \gamma_1(1) = \gamma_0,$$

i.e., $G \subset \Gamma$. Therefore, for $\gamma \in G$ it holds that

$$0 < \omega \leqq 1 < \frac{2}{1+\tilde{b}}, \qquad \omega' \in \left(1, \frac{2}{1+\tilde{q}}\right).$$

In the third case, we must prove only that

$$b_i < \frac{b_i\omega}{2-\omega} < \gamma_0 \quad \text{for } \omega \in \left(1, \frac{2}{1+\tilde{b}}\right).$$

It is easy to verify that $1 < \frac{\omega}{2-\omega}$ for $\omega > 1$, from which follows the first inequality.

From (3.7), we obtain

$$\omega < \frac{2}{1 + (b_i/\gamma_0)},$$

and thus

$$\frac{b_i\omega}{2-\omega} < \gamma_0,$$

i.e., $G \subset \Gamma$. Now, if

$$\gamma \in G \quad \text{and} \quad \omega \in \left(1, \frac{2}{1+\tilde{b}}\right),$$

(3.5) follows.

In the last case we have to prove only that

$$\frac{b_i \omega}{2-\omega} < \gamma_1(\omega') \leqq \gamma_0,$$

since we know that in this case it holds that

$$b_i < \frac{b_i \omega}{2-\omega} \quad \text{and} \quad \gamma_1(\omega) \leqq \gamma_0.$$

The above inequality follows from $\omega' < \phi(\omega)$, which we can obtain after simple manipulations. Now, for each $\gamma \in G$, (3.5) is satisfied. □

4. Numerical examples. In this section, we present some matrices in order to show that each of our classes is not contained in the others. We also give intervals of convergence for these matrices.

Example 1. Let

$$A_1 = \begin{bmatrix} 1 & .5 & .4 \\ 1.2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

It is easy to see that A_1 does not belong to the classes C_0, C_2, and C_3, but belongs to the class C_1 for $\frac{2}{3} < \alpha < \frac{5}{7}$. From [10], we have the following intervals of convergence for the MSOR method for $\alpha = 0.7$, when it approximates the solution of $A_2 x = b$:

$$0 < w \leqq 1 \quad \text{and} \quad \frac{15}{16} w < w' < \frac{185}{184},$$

or

$$1 < \omega < \frac{164}{163} \quad \text{and} \quad \frac{15}{16}\omega < \omega' < \min\left\{\frac{200-15\omega}{184}, \frac{200-163\omega}{36}\right\}.$$

Example 2. Let

$$A_2 = \begin{bmatrix} 1 & .1 & .9 \\ 1.1 & 1 & 0 \\ .9 & 0 & 1 \end{bmatrix}.$$

For this matrix we can see that A_2 belongs to the class C_2, but does not belong to the classes C_0, C_1, and C_3. From Theorem 3.4, we have the following intervals of convergence for the MSOR method:

$$\omega \in \left(0, \frac{2}{2.1}\right) \quad \text{and} \quad \omega' \in \left(0, \frac{2-w}{1+0.05w}\right).$$

Example 3. Let

$$A_3 = \begin{bmatrix} 1 & 0 & 1.3 & 0 & 0 \\ 0 & 1 & .5 & .1 & 0 \\ .3 & .5 & 1 & 0 & 0 \\ .2 & .1 & 0 & 1 & 0 \\ .5 & .1 & 0 & 0 & 1 \end{bmatrix}.$$

This matrix belongs to the class C_3, but does not belong to C_0, C_1, or C_2. Following Theorem 3.8, we obtain the intervals of convergence for the MSOR method given by

$$\omega \in (0, 1] \quad \text{and} \quad \omega' \in \left(0, \frac{2}{1.89}\right),$$

or

$$\omega \in \left(1, \frac{2}{1.975}\right) \quad \text{and} \quad \omega' \in \left(0, \frac{2-\omega}{1.5 - .555\omega}\right),$$

with

$$i = 1, \quad \gamma_0 = \tfrac{5}{3}, \quad b_1 = 1.3, \quad b' = .6.$$

Acknowledgment. The authors would like to thank the referee for his valuable suggestions.

REFERENCES

[1] R. Beauwens, *Semistrict diagonal dominance*, SIAM J. Numer. Anal., 13 (1976), pp. 109–112.

[2] Lj. Cvetković and D. Herceg, *Some results on M- and H-matrices*, Univ. u Novom Sadu, Zb. Rad. Prirod., Mat. Fak. Ser. Mat., 17 (1987), pp. 121–129.

[3] ———, *Improvement of the area of convergence of the* AOR *method*, Anal. Numer. Theory Approx., 16 (1987), pp. 109–115.

[4] ———, *The* AOR *method for solving linear interval equations*, Computing, 41 (1989), pp. 359–364.

[5] ———, *Convergence theory for the* AOR *method*, J. Comput. Math., 8 (1990), pp. 675–693.

[6] D. Herceg, *Remarks on different splittings and associated generalized linear methods*, Univ. u Novom Sadu, Zb. Rad. Prirod., Mat. Fak. Ser. Mat., 13 (1983), pp. 177–186.

[7] ———, *Über die Konvergenz des* AOR-*Verfahrens*, Z. Angew. Math. Mech., 65 (1985), pp. 378–379.

[8] ———, *On a modification of nonlinear overrelaxation methods*, Univ. u Novom Sadu, Zb. Rad. Prirod., Mat. Fak. Ser. Mat., 19 (1989).

[9] D. Herceg and Lj. Cvetković, *On the extrapolation method and* USA *algorithm*, J. Econom. Dynamics Control, 13 (1988), pp. 301–311.

[10] D. Herceg, M. M. Martins, and M. E. Trigo, *Extended convergence area for the* MSOR *method*, J. Comp. Appl. Math., 33 (1990), pp. 123–132.

[11] K. R. James and A. Riha, *Convergence criteria for successive overrelaxation*, SIAM J. Numer. Anal., 12 (1975), pp. 137–143.

[12] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, MA, 1964.

[13] M. M. Martins, *On the convergence of the modified overrelaxation method*, Linear Algebra Appl., 81 (1986), pp. 55–73.

[14] ———, *A note on the convergence of the* MSOR *method* II, Linear Algebra Appl., 141 (1990), pp. 223–226.

[15] R. S. Varga, *On recurring theorems and diagonal dominance*, Linear Algebra Appl., 13 (1976), pp. 1–9.

[16] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

MORE MATRIX FORMS OF THE ARITHMETIC-GEOMETRIC MEAN INEQUALITY*

RAJENDRA BHATIA† AND CHANDLER DAVIS‡

Dedicated to T. Ando on his 60th birthday

Abstract. For arbitrary $n \times n$ matrices A, B, X, and for every unitarily invariant norm, it is proved that $2|||A^*XB||| \leqq |||AA^*X + XBB^*|||$.

Key words. geometric mean, singular values, unitarily invariant norm

AMS(MOS) subject classifications. 15A42, 15A60, 47A30, 47B05, 47B10

In an earlier paper [3] it was proved that, for arbitrary $n \times n$ matrices A and B,

$$2s_j(A^*B) \leqq s_j(AA^* + BB^*), \qquad j = 1, 2, \ldots, n, \tag{1}$$

where s_j are the singular values in decreasing order. This means, in particular, that

$$2|||A^*B||| \leqq |||AA^* + BB^*||| \tag{2}$$

for every unitarily invariant norm.

Our main result is a considerable strengthening of the latter inequality.

THEOREM 1. *For arbitrary $n \times n$ matrices A, B, X,*

$$2|||A^*XB||| \leqq |||AA^*X + XBB^*||| \tag{3}$$

for every unitarily invariant norm.

(The corresponding strengthening of statement (1) is easily seen *not* to hold, even for positive-definite 2×2 matrices.)

An incidental benefit from our proof is the insight it may afford into the order of the factors in (1) and (2). That the order is critical was shown in [3]; Theorem 1 may make it look less surprising.

For the special case of the bound norm, (3) was discovered earlier by McIntosh [6] with a different proof (and a different motivation). Fuad Kittaneh found the Hilbert-Schmidt case of (3), with conditions for equality, before we began our work. He also made subsequent contributions, which are being published elsewhere.

COROLLARY. *For positive semidefinite matrices A, C, and for any unitarily invariant norm,*

$$|||AC^2A||| \leqq |||A^2C^2|||.$$

(The case of the p-norm, with p a power of 2, is a known inequality [7].) This follows from Theorem 1 by

$$|||AC^2A||| \leqq \tfrac{1}{2}|||A^2C^2 + C^2A^2||| \leqq |||A^2C^2|||.$$

The theorem can be strengthened in case A and B are positive semidefinite.

* Received by the editors April 11, 1990; accepted for publication (in revised form) May 9, 1991.

† Indian Statistical Institute, 7, SJS Sansanwal Marg, New Delhi 110 016, India (vikram!isid!rbh@ncst.ernet.in).

‡ Department of Mathematics, University of Toronto, Toronto M5S 1A1, Canada (davis@math.toronto.edu).

THEOREM 2. *For $n \times n$ matrices A, B, X with $A \geqq 0$, $B \geqq 0$, and for any unitarily invariant norm, the real function*

$$f(p) = |||A^{1+p}XB^{1-p} + A^{1-p}XB^{1+p}||| \tag{4}$$

is convex on $[-1, 1]$ and takes its minimum at $p = 0$.

Another result, akin to Theorem 1 but with a separate proof, is proved in Theorem 3.

THEOREM 3. *Fix an $n \times n$ matrix G, and any unitarily invariant norm. Among choices of F and H which make the matrix*

$$\begin{bmatrix} F & G \\ G^* & H \end{bmatrix}$$

positive semidefinite, that which minimizes its norm is $F = |G^*|$, $H = |G|$. (As is customary, $|G|$ denotes $(G^*G)^{1/2}$.)

This is an especially simple example of efficient completion of a partial matrix [4], [5], in this case of

$$\begin{bmatrix} ? & G \\ G^* & ? \end{bmatrix}.$$

It differs from many other situations in that here the same completion is efficient for many norms; contrast [4].

Proof of Theorem 1. The main features already appear in a particular case.

Case 1. $B = A \geqq 0$, $X \geqq 0$. First take the bound norm $\| \;\|$. It is to be proved that

$$2\|AXA\| \leqq \|A^2X + XA^2\|.$$

With both sides depending continuously upon A, it is enough to treat the case where A is invertible. Then, introducing the notation $T = A^2X = A(AXA)A^{-1}$, we note that $\sigma(T) = \sigma(T^*) = \sigma(AXA)$; the norm of $2AXA$ is to be compared with that of $T + T^*$.

Now, AXA is positive, so its norm is the maximum among its eigenvalues, say λ_1. In light of $\lambda_1 \in \sigma(T) \subseteq W(T)$, the numerical range of T, there is some unit vector x with $x^*Tx = \lambda_1$ (it could be an eigenvector belonging to eigenvalue λ_1, but it need not be). For the same x, $x^*T^*x = \bar{\lambda}_1 = \lambda_1$, so that

$$2\|AXA\| = 2\lambda_1 = x^*(T + T^*)x \leqq \|T + T^*\|,$$

as desired.

We go from this to the assertion for arbitrary unitarily invariant norms by a familiar procedure. First, it is enough to obtain the conclusion for the kth Ky Fan norm $\| \;\|_k$ for arbitrary $k = 1, 2, \ldots, n$; see [2, § 7]. We have now done it for $\| \;\| = \| \;\|_1$.

By $\Lambda^k A$ denote the operator $A \otimes \cdots \otimes A$ (k factors) restricted to its invariant subspace $\Lambda^k(C^n)$ consisting of all skew k-tensors; by $T^{[k]}$ denote the operator $T \otimes 1 \otimes \cdots \otimes 1 + 1 \otimes T \otimes \cdots \otimes 1 + \cdots + 1 \otimes \cdots \otimes 1 \otimes T$ restricted to its invariant subspace $\Lambda^k(C^n)$, and similarly $(AXA)^{[k]}$; see [2, § 6]. Now $T^{[k]} = (\Lambda^k A)(AXA)^{[k]}(\Lambda^k A)^{-1}$, so $\sigma(T^{[k]}) = \sigma((AXA)^{[k]})$ and the reasoning used above for AXA and T can be invoked. Remember that the eigenvalues of $(AXA)^{[k]}$ are obtained from the eigenvalues $\lambda_1 \geqq \lambda_2 \geqq \cdots$ of AXA as all sums $\lambda_{i_1} + \cdots + \lambda_{i_k}$ formed using distinct indices $i_1, \ldots, i_k$, and remember that $(AXA)^{[k]} \geqq 0$. So the conclusion is that for some unit vector y of $\Lambda^k(C^n)$,

$$\begin{aligned} 2\|AXA\|_k &= 2(\lambda_1 + \cdots + \lambda_k) \\ &= 2\|(AXA)^{[k]}\| = y^*(T^{[k]} + (T^{[k]})^*)y \leqq \|(T + T^*)^{[k]}\|. \end{aligned} \tag{5}$$

On the right we have the norm of a Hermitian matrix, in particular, the modulus of one of its eigenvalues. But we know these are sums of eigenvalues of $T + T^*$; hence, the right-hand member of (5) is less than or equal to the sum of the top k singular values of $T + T^*$, viz., $\|T + T^*\|_k$. Case 1 is complete.

We turn to the general case. This is proved by reducing it, not to Case 1, but to a form where the same ideas can be tricked into serving.

First, consider the polar resolutions $A = A_1 V$, $B = B_1 W$ (with A_1 and B_1 positive semidefinite and V and W unitary). On the right-hand side in (3), we are considering $AA^*X + XBB^* = A_1^2 X + XB_1^2$, while on the left we are considering $A^*XB = V^*A_1XB_1W$, which has the same norms as A_1XB_1. Therefore, the general case will be established if we prove the result for $A \geqq 0$, $B \geqq 0$, X arbitrary. As before, we are also free to assume that A and B are invertible.

We begin by defining the operators

$$A_0 = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \geqq 0, \qquad X_0 = \begin{bmatrix} 0 & X \\ X^* & 0 \end{bmatrix}.$$

With this definition,

$$(6) \qquad \begin{aligned} A_0X_0A_0 &= \begin{bmatrix} 0 & AXB \\ BX^*A & 0 \end{bmatrix}, \\ A_0^2X_0 + X_0A_0^2 &= \begin{bmatrix} 0 & A^2X + XB^2 \\ B^2X^* + X^*A^2 & 0 \end{bmatrix}. \end{aligned}$$

These matrices have eigenvalues that come in pairs. Thus we may write the eigenvalues of $A_0X_0A_0$ as $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n \geqq -\lambda_n \geqq \cdots \geqq -\lambda_1$, where the λ_j are the singular values of AXB, and the eigenvalues of $A_0X_0^2 + X_0A_0^2$ as $\pm\mu_j$, where $\mu_1 \geqq \cdots \geqq \mu_n$ are the singular values of $A^2X + XB^2$.

To adapt the ideas of Case 1, we will estimate the quantity $\|(A_0X_0A_0)^{[k]}\|$ $(1 \leqq k \leqq n)$. As the norm of a Hermitian operator, it is given by an eigenvalue, hence by a sum of some $\pm\lambda_j$. Choosing these to maximize the modulus of the sum requires choosing all to have the same sign; that is, $\|(A_0X_0A_0)^{[k]}\| = \lambda_1 + \cdots + \lambda_k = \|AXB\|_k$. Similarly for the other matrix in (6).

Now let $T = A_0^2X_0 = A_0(A_0X_0A_0)A_0^{-1}$. Its spectrum is the same as that of $A_0X_0A_0$. We must consider the second matrix in (6), which is $T + T^*$. Reasoning as in Case 1, there is some unit vector $z \in \Lambda^k(C^{2n})$ such that

$$\begin{aligned} 2\|AXB\|_k &= 2(\lambda_1 + \cdots + \lambda_k) = z^*(T + T^*)^{[k]}z \\ &\leqq \|(T + T^*)^{[k]}\| = \mu_1 + \cdots + \mu_k = \|A^2X + XB^2\|_k. \end{aligned}$$

That is, $2|||AXB||| \leqq |||A^2X + XB^2|||$ holds for all the Ky Fan norms, whence it holds for all unitarily invariant norms. □

We have an alternative proof of Theorem 1 which is in some respects preferable. In particular, the recourse to tensor algebra can be avoided. We sketch the key part, the proof that for $A = A^*$ and $X = X^*$, we have that

$$(7) \qquad 2\sum_{j=1}^{k} s_j(AXA) \leqq \sum_{j=1}^{k} s_j(A^2X + XA^2).$$

Assume without loss of generality that A is invertible. Then, letting $T = A^2X$ so that $T^* = XA^2$, we have that $T = A(AXA)A^{-1}$ and $T^* = A^{-1}(AXA)A$, whence the three matrices AXA, T, and T^* all have the same spectrum. Denote these eigenvalues by λ_1, $\lambda_2, \ldots$, in order of decreasing $|\lambda_j|$. Because AXA is Hermitian, its singular values are $s_j = |\lambda_j|$, so that the left-hand member of (7) is just

$$2\sum_{j=1}^{k} |\lambda_j|.$$

When T is put in Schur upper-triangular form with diagonal $(\lambda_1, \lambda_2, \ldots)$, we are referring it to an orthonormal basis, $(e_1, e_2, \ldots, e_n)$, for which

$$\lambda_j = e_j^* Te_j = e_j^* T^* e_j.$$

(T^* takes lower-triangular form relative to this basis.) Therefore

$$2\sum_{j=1}^{k} |\lambda_j| = \sum_{j=1}^{k} |e_j^*(T+T^*)e_j| \leqq \sum_{j=1}^{k} s_j(T+T^*),$$

the inequality by Ky Fan's variational criterion [2]. This establishes (7).

Proof of Theorem 2. Again, by continuity, we may assume that A and B are invertible. Since f is plainly continuous and $f(-p) = f(p)$, both conclusions of the theorem will follow if we now show that $2f(p) \leqq f(p+q) + f(p-q)$ when $p \pm q$ both lie in $[-1, 1]$.

Consider the mapping on positive matrices defined by

$$2\mathcal{M}_p(Y) = A^pYB^{-p} + A^{-p}YB^p.$$

Theorem 1 tells us that $|||\mathcal{M}_q(Y)||| \geqq |||Y|||$. Apply this to $Y = \mathcal{M}_p(AXB)$, using the identity $2\mathcal{M}_q(\mathcal{M}_p(Y)) = \mathcal{M}_{p+q}(Y) + \mathcal{M}_{p-q}(Y)$, and the result is

$$\begin{aligned}2f(p) &= 4|||\mathcal{M}_p(AXB)||| \leqq 4|||\mathcal{M}_q(\mathcal{M}_p(AXB))|||\\ &\leqq 2|||\mathcal{M}_{p+q}(AXB)||| + 2|||\mathcal{M}_{p-q}(AXB)||| = f(p+q) + f(p-q). \qquad \square\end{aligned}$$

Proof of Theorem 3. Consider the polar resolution $G = UK$ with U unitary. Then the matrix under study,

$$\begin{bmatrix} F & G \\ G^* & H \end{bmatrix},$$

has the same norms as its unitary transform

$$\begin{bmatrix} U^* & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} F & G \\ G^* & H \end{bmatrix}\begin{bmatrix} U & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} L & K \\ K & H \end{bmatrix}.$$

This, in turn, has the same norms as its unitary transform

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} L & K \\ K & H \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} H & K \\ K & L \end{bmatrix}.$$

Without increasing any norm, we can go from one of these to their mean. $[\begin{smallmatrix} M & K \\ K & M \end{smallmatrix}]$ where $M = \frac{1}{2}(H + L)$. This positive-semidefinite matrix has the same norm as its unitary transform (by the unitary $2^{-1/2}[\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}]$)

$$\begin{bmatrix} M+K & 0 \\ 0 & M-K \end{bmatrix}.$$

This evidently satisfies

$$\begin{bmatrix} M+K & 0 \\ 0 & M-K \end{bmatrix} \geqq \begin{bmatrix} 2K & 0 \\ 0 & 0 \end{bmatrix} \geqq 0,$$

which is known to imply

$$\left|\left|\left| \begin{bmatrix} M+K & 0 \\ 0 & M-K \end{bmatrix} \right|\right|\right| \geqq \left|\left|\left| \begin{bmatrix} 2K & 0 \\ 0 & 0 \end{bmatrix} \right|\right|\right|.$$

But this comparison matrix $[\begin{smallmatrix} 2K & 0 \\ 0 & 0 \end{smallmatrix}]$ is just what matrix

$$\begin{bmatrix} M+K & 0 \\ 0 & M-K \end{bmatrix}$$

reduces to when, in particular, $F = |G^*|, H = |G|, H = L = K = M$. Therefore, undoing all the unitary similarities leads to the inequality announced. □

The reason we see this result as related to Theorem 1 may be seen in the special case in which $G \geqq 0$ is invertible. Those pairs (F, H) that retain positivity of $[\begin{smallmatrix} F & G \\ G & H \end{smallmatrix}]$ while minimizing the rank are those obtained by $H = GF^{-1}G$. Such pairs all have the property that $G = F \# H$, the geometric mean of F and H [1]. The most efficient of them, in some sense, should be (G, G). Theorem 3 gives one sense in which this holds. Another is that $F + GF^{-1}G \geqq 2G$, a familiar elementary computation. (Taking traces in this inequality gives Theorem 3 in the special case of the trace norm.)

Acknowledgments. Davis thanks the Indian Statistical Institute and the Natural Science and Engineering Research Council (Canada) for making possible a visit during which this work was done. We thank two referees for correcting our faulty version of Theorem 2 and for contributing the corollary.

REFERENCES

[1] T. Ando, *Concavity of certain maps on positive definite matrices and applications to Hadamard products*, Linear Algebra Appl., 26 (1979), pp. 203–241.

[2] R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, Longman, Essex and Wiley, New York, 1987.

[3] R. Bhatia and F. Kittaneh, *On the singular values of a product of operators*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 272–277.

[4] C. Davis, *An extremal problem for extensions of a sesquilinear form*, Linear Algebra Appl., 13 (1976), pp. 91–102.

[5] C. Davis, W. M. Kahan, and H. F. Weinberger, *Norm-preserving dilations and their applications to optimal error bounds*, SIAM J. Numer. Anal., 19 (1982), pp. 445–469.

[6] A. McIntosh, *Heinz inequalities and perturbation of spectral families*, Macquarie Mathematics Reports 79-0006, 1979.

[7] C. J. Thompson, *Inequality with applications in statistical mechanics*, J. Math. Phys., 6 (1965), pp. 1812, 1813.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 137–145, January 1993

SOME RESULTS ON HOMOGENEOUS MATRIX EQUATIONS*

D. VON ROSEN†

Abstract. New representations of general solutions to the matrix equations $AXB = 0$ and $A_1XB_1 = 0$, $A_2XB_2 = 0$, based on vector space decompositions, are given. In a certain sense the solutions are natural and show an easily interpretable structure. Furthermore, the solutions enable the finding of solutions to $A_1X_1B_1 + A_2X_2B_2 = 0$ and $A_1XB_1 = 0$, $A_2XB_2 = 0$, $A_3XB_3 = 0$ where for the latter equations no solutions have hitherto been derived. Some other extensions are also considered.

Key words. matrix equations, decomposing vector spaces, tensor product, growth curve model

AMS(MOS) subject classifications. 15A06, 15A03

1. Introduction. Matrix equations play an important role in many fields and this paper deals with homogeneous matrix equations. Although several of the equations discussed here have been treated by other authors, our solutions will, in our opinion, be more natural. For instance, consider the following equation in X, $AXB = 0$, where the included matrices are supposed to be of proper order. The well known and most commonly applied solution is given by $X = Z - A^-AZBB^-$, where Z is an arbitrary matrix and $^-$ signifies an arbitrary g-inverse in the sense of $GG^-G = G$. Note that Z has the same size as X, which is unnatural since all solutions to $AXB = 0$ can be generated by a fewer number of arbitrary elements than the size of X. Keeping this in mind, we obtain here an alternative representation of the solution.

The aim is to present an approach that rests on the idea of decomposing vector spaces. It is easy to apply the method to $AXB = 0$ and $A_iXB_i = 0$, $i = 1, 2$, and the solutions obtained in that manner are fairly useful. In this paper they are utilized when solving $A_1X_1B_1 + A_2X_2B_2 = 0$ and $A_iXB_i = 0$, $i = 1, 2, 3$. Furthermore, $A_iXB_i = 0$, $i = 1, 2, \ldots, s$ is discussed under certain constraints on the B's.

The matrices in this paper consist of real elements, $'$ stands for the transpose, and $R(\cdot)$ and $O(\cdot)$ denote the range space and orthogonal complement. A matrix A^O is any matrix satisfying $R(A^O) = O(A)$. The symbol $\oplus$ indicates that the sum of vector spaces consists of disjoint subspaces and $\boxplus$ means that the subspaces, additionally, are orthogonal. Moreover, vec $(\cdot)$ signifies the vec-operator defined by $R^{n\times m} \rightarrow R^{nm}$, $xy' \rightarrow y \otimes x$ where $x: n \times 1$, $y: m \times 1$, and $\otimes$ stands for the right Kronecker product. Some further definitions of minor importance will be given in the text. Finally, we note that the results can be easily extended to comprise complex matrices and the necessary alterations are indicated after Theorem 1 and Theorem 2.

In order to characterize classes of g-inverses, Ben-Israel and Greville [3, p. 77] discussed the equation $AXA = 0$. They obtained a solution $X = FZ_1 + BZ_2K$ where Z_1 and Z_2 are arbitrary matrices and F, B, and K are matrices whose columns generate certain vector spaces. Our solution to $AXA = 0$ is almost identical to Ben-Israel and Greville's, but the method of obtaining it differs.

Another application where the solutions in this paper appear to be advantageous is when finding estimators for parameters under linear restrictions in a multivariate linear statistical model, namely, the well-known Growth Curve model (Potthoff and Roy [8]).

* Received by the editors December 4, 1989; accepted for publication (in revised form) June 25, 1991.

† Department of Mathematics, Uppsala University, Box 480, S-751 06 Uppsala, Sweden (math@tdb.uu.se).

We end the introduction by giving some details on this topic and making some comparisons between the solution $X = Z - A^{-}AZBB^{-}$ and Theorem 1 (i) in this paper.

The Growth Curve model is defined by the equation $X = ABC + W$, where A and C are known matrices and the columns of W are independent normally distributed with mean zero and a covariance matrix Σ. The matrices B and Σ consist of the unknown parameters, which are to be estimated. We suppose that $DBE = 0$ holds where D and E are known matrices. By a reparametrization, utilizing Theorem 1 (i) in this paper, it follows that we can equivalently study a model defined by

$$X = A(D')^{O}\theta_1 C + AD'\theta_2 E^{O\prime}C + W, \tag{1.1}$$

where θ_1 and θ_2 now are new parameter matrices. Instead of (1.1), if using the "classical" solution for $DBE = 0$, $B = \theta - D^{-}D\theta EE^{-}$ (θ is a new parameter matrix), the model

$$X = A\theta C - AD^{-}D\theta EE^{-}C + W \tag{1.2}$$

is obtained. We are going to make some comparisons between the models in (1.1) and (1.2) leading to the conclusion that (1.1) is easier to handle than (1.2). The likelihood equations for θ_1, θ_2, and Σ in (1.1), and θ and Σ in (1.2), are, respectively, given by (n is the number of columns in X)

$$\begin{gathered}(A(D')^{O})'\Sigma^{-1}(X - A(D')^{O}\theta_1 C - AD'\theta_2 E^{O\prime}C)C' = 0,\\ (AD')'\Sigma^{-1}(X - A(D')^{O}\theta_1 C - AD'\theta_2 E^{O\prime}C)(E^{O\prime}C)' = 0,\\ n\Sigma = (X - A(D')^{O}\theta_1 C - AD'\theta_2 E^{O\prime}C)(X - A(D')^{O}\theta_1 C - AD'\theta_2 E^{O\prime}C)',\end{gathered} \tag{1.3}$$

and

$$\begin{gathered}A'\Sigma^{-1}(X - A\theta C + AD^{-}D\theta EE^{-}C)C'\\ - D'D^{-\prime}A'\Sigma^{-1}(X - A\theta C + AD^{-}D\theta EE^{-}C)C'E^{-\prime}E' = 0,\\ n\Sigma = (X - A\theta C + AD^{-}D\theta EE^{-}C)(X - A\theta C + AD^{-}D\theta EE^{-}C)'.\end{gathered} \tag{1.4}$$

Both (1.3) and (1.4) are nonlinear equations. To find explicit expressions for the parameters in (1.4) seems rather difficult and is still an open problem. However, for (1.3) it is feasible, and we briefly show how to solve this nonlinear equation system. This may be of interest to some readers since solving, in fact, is fairly easy and can be useful in other situations. Further details are presented by von Rosen [10]. Set (I denotes the identity matrix)

$$\begin{aligned}A_1 &= A(D')^{O}, \quad A_2 = AD', \quad C_1 = C, \quad C_2 = E^{O\prime}C,\\ V_1 &= XC_1'(C_1C_1')^{-}C_1 - \sum_{i=1}^{2} A_i\theta_i C_i,\\ V_2 &= XC_2'(C_2C_2')^{-}C_2 - A_2\theta_2 C_2,\\ S_1 &= X(I - C_1'(C_1C_1')^{-}C_1)X',\\ P_2 &= I - A_1(A_1'S_1^{-1}A_1)^{-}A_1'S_1^{-1},\\ S_2 &= S_1 + P_2X(C_1'(C_1C_1')^{-}C_1 - C_2'(C_2C_2')^{-}C_2)X'P_2'.\end{aligned}$$

The likelihood equations in (1.3) can be written

$$A'_r\Sigma^{-1}\left(X - \sum_{i=1}^{2} A_i\theta_iC_i\right)C'_r = 0, \qquad r = 1, 2, \tag{1.5}$$

$$n\Sigma = \left(X - \sum_{i=1}^{2} A_i\theta_iC_i\right)\left(X - \sum_{i=1}^{2} A_i\theta_iC_i\right)'. \tag{1.6}$$

Starting with $r = 1$ in (1.5) after some manipulations we have the equations

$$A'_1\Sigma^{-1}V_1C'_1 = 0, \tag{1.7}$$

$$n\Sigma = S_1 + V_1V'_1. \tag{1.8}$$

Now the key to solving the likelihood equations is to use the well-known relation $(S + VV')^{-1}V = S^{-1}VH$, where H is some positive definite matrix, and obtain, by inserting (1.8) into (1.7),

$$A'_1S_1^{-1}V_1HC'_1 = 0. \tag{1.9}$$

However, since by definition of C'_2 and C'_1, $R(C'_2) \subseteq R(C'_1)$, and since H is positive definite, we have $R(C_1HC'_2) = R(C_1)$, implying by the definition of V_1 that (1.9) is independent of H and therefore is a consistent linear equation in θ_1 with an easily obtainable solution, e.g., use Theorem 1(i) in this paper. Inserting the solution for θ_1 into (1.5) for $r = 2$ and in (1.8), and doing some calculations, we obtain

$$A'_2\Sigma^{-1}P_2(X - A_2\theta_2C_2)C'_2 = 0, \tag{1.10}$$

$$n\Sigma = S_2 + P_2(X - A_2\theta_2C_2)(X - A_2\theta_2C_2)'P'_2. \tag{1.11}$$

Using the same ideas as above leads to the equation in θ_2:

$$A'_2S_2^{-1}P_2V_2HC'_2 = 0, \tag{1.12}$$

where H, although differing from H in (1.9), is positive definite. Thus (1.12) is a linear equation in θ_2, and inserting the solution for θ_2 into (1.11) and into the solution for θ_1 obtained with the help of (1.9), enables all parameters to be estimated. In summary, we have the following maximum likelihood estimators (note that $P'_2S_2^{-1}P_2 = S_2^{-1}P_2$):

$$\hat{\theta}_1 = (A'_1S_1^{-1}A_1)^-A'_1S_1^{-1}(X - A_2\hat{\theta}_2C_2)C'_1(C_1C'_1)^- + (A'_1)^OZ_1 + A'_1Z_2C_1^{O\prime},$$

$$\hat{\theta}_2 = (A'_2P'_2S_2^{-1}P_2A_2)^-A'_2P'_2S_2^{-1}XC'_2(C_2C'_2)^- + (A'_2P'_2)^OZ_3 + A'_2P'_2Z_4C_2^{O\prime},$$

$$n\hat{\Sigma} = \left(X - \sum_{i=1}^{2} A_i\hat{\theta}_iC_i\right)\left(X - \sum_{i=1}^{2} A_i\hat{\theta}_iC_i\right)',$$

where the Z's are arbitrary matrices.

One reason why it is easier to discuss (1.3) instead of (1.4) is that in (1.1) we obtain a correct parametrization, whereas in (1.2) the model is overparametrized. Hence, in comparison with (1.2), we have indicated some advantages of (1.1) which, in turn, was obtained with the help of the results on homogeneous matrix equations given in the next section.

2. Main results. The object in this section is to find solutions to the following equations in X:

$$AXB = 0, \tag{2.1}$$

where $A: p \times q$, $B: k \times n$, and

$$A_1XB_1 = 0, \tag{2.2}$$

$$A_2XB_2 = 0, \tag{2.3}$$

where A_i: $p_i \times q$ and B_i: $k \times n_i$, $i = 1, 2$. Solutions to (2.1) can be found in many textbooks and one reference is Rao and Mitra [9]. A solution to (2.2) and (2.3) was first obtained by Mitra [5], who went on to discuss it further in [6]. With the help of vector space decomposition, Banken [2] treated a special case of (2.2) and (2.3), namely, when $R(A_2') \subseteq R(A_1')$. Here we will see that (2.2) and (2.3) can be handled in a manner similar to the equation given by (2.1). The approach in this section is designed for dealing with homogeneous equations, whereas Mitra's intention [5], [6] was to discuss inhomogeneous equations. Therefore, this paper, in the homogeneous case, gives more detailed information about the solutions than the approaches taken by Mitra.

THEOREM 1. (i) *A representation of the general solution to* (2.1) *is*

$$X = (A')^O Z_1 B' + A' Z_2 B^{O'} + (A')^O Z_3 B^{O'}$$

or

$$X = (A')^O Z_1 + A' Z_2 B^{O'}$$

or

$$X = Z_1 B^{O'} + (A')^O Z_2 B',$$

where the Z's are arbitrary matrices of proper order.

(ii) *A representation of the general solution to* (2.2) *and* (2.3) *is*

$$\begin{aligned} X = {} & T_3 Z_1 S_1' + T_4 Z_2 S_1' + T_4 Z_3 S_2' + T_1 Z_4 S_3' + T_4 Z_5 S_3' \\ & + T_1 Z_6 S_4' + T_2 Z_7 S_4' + T_3 Z_8 S_4' + T_4 Z_9 S_4' \end{aligned}$$

or

$$X = (A_2')^O Z_1 S_1' + (A_1' : A_2')^O Z_2 S_2' + (A_1')^O Z_3 S_3' + Z_4 S_4'$$

or

$$X = T_1 Z_1 B_2^{O'} + T_2 Z_2 (B_1 : B_2)^{O'} + T_3 Z_3 B_1^{O'} + T_4 Z_4,$$

where the Z's are arbitrary matrices of proper order and $S_1 - S_4$, $T_1 - T_4$ are any matrices satisfying (*$(A : B)$ means the partitioned matrix of A and B*)

$$\begin{aligned} R(S_1) &= R(B_1 : B_2) \cap O(B_1), & R(T_1) &= R(A_1' : A_2') \cap O(A_1'), \\ R(S_2) &= R(B_1) \cap R(B_2), & R(T_2) &= R(A_1') \cap R(A_2'), \\ R(S_3) &= R(B_1 : B_2) \cap O(B_2), & R(T_3) &= R(A_1' : A_2') \cap O(A_2'), \\ R(S_4) &= O(B_1 : B_2), & R(T_4) &= O(A_1' : A_2'). \end{aligned}$$

Proof. Since $AXB = 0$ is equivalent to $(B' \otimes A)$ vec $(X) = 0$, we must consider $O(B \otimes A')$, which equals

$$\begin{aligned} O(B \otimes A') &= R(B \otimes (A')^O) \boxplus R(B^O \otimes A') \boxplus R(B^O \otimes (A')^O) \\ &= R(B \otimes (A')^O) \boxplus R(B^O \otimes I) \\ &= R(I \otimes (A')^O) \boxplus R(B^O \otimes A'), \end{aligned}$$

showing (i). Statement (ii) follows since

$$
\begin{aligned}
O((B_1 \otimes A_1') : (B_2 \otimes A_2')) &= R((S_1 \otimes T_3) : (S_1 \otimes T_4) : (S_2 \otimes T_4) : (S_3 \otimes T_1) : \\
&\qquad (S_3 \otimes T_4) : (S_4 \otimes T_1) : (S_4 \otimes T_2) : (S_4 \otimes T_3) : (S_4 \otimes T_4)) \\
&= (R(S_1 \otimes (A_2')^o) \oplus R(S_3 \otimes (A_1')^o)) \boxplus R(S_2 \otimes (A_1' : A_2')^o) \boxplus R(S_4 \otimes I) \\
&= (R(B_2^o \otimes T_1) \oplus R(B_1^o \otimes T_3)) \boxplus R((B_1 : B_2)^o \otimes T_2) \boxplus R(I \otimes T_4).
\end{aligned}
$$

One way of verifying these relations is by applying the orthomodular law (see, e.g., Birkhoff [4]) to tensor products of finite-dimensional vector spaces. Some details of doing this can be found in Nordström and von Rosen [7]. Another way is to note that the right-hand side is a subspace of $O((B_1 \otimes A_1') : (B_2 \otimes A_2'))$ and thereafter show that both sides have the same dimension. □

Remark. When dealing with complex matrices, instead of $O(B \otimes A')$ and $O((B_1 \otimes A_1') : (B_2 \otimes A_2'))$, we must consider $O(\bar{B} \times \bar{A}')$ and $O((\bar{B}_1 \otimes \bar{A}_1') : (\bar{B}_2 \otimes \bar{A}_2'))$, where the overbar represents the conjugate. Moreover, we note that it is fairly easy to find representations for the S's and T's in part (ii) of the theorem.

In (i) and (ii) of Theorem 1 we can replace the matrices with full-rank matrices as long as the spaces are unaffected. This implies that the number of "free" elements can be chosen to be in direct correspondence with the restrictions imposed by the equations. As already mentioned, we mean that this property implies that the solutions presented in the theorem are more natural than those commonly applied. Another advantage with the above solutions is that it is possible to obtain conditions on matrices K and L such that KXL, under the restrictions in Theorem 1, belongs to a certain subspace. A special case is, of course, $KXL = 0$. From the theorem it is also easy to get conditions on the matrices so that other solutions than the trivial one $X = 0$ are obtained. Furthermore, in a statistical context the Z's in the theorem may have natural interpretations. For example, models of the form (1.1) are discussed by Verbyla and Venables [11].

The orthogonal complement used in the proof of Theorem 1(ii) rests on a decomposition of the entire space based on

$$
\begin{aligned}
R(I) &= R(S_1) + R(S_2) + R(S_3) + R(S_4), \\
R(I) &= R(T_1) + R(T_2) + R(T_3) + R(T_4).
\end{aligned}
$$

However, if $R(B_1) \subseteq R(B_2)$ holds in Theorem 1(ii), we obtain solutions that do not immediately follow from the theorem. Under the condition $R(B_1) \subseteq R(B_2)$, it is straightforward to show that

$$O((B_1 \otimes A_1') : (B_2 \otimes A_2')) = R(B_1^o \otimes T_3) \boxplus R(B_2^o \otimes A_2') \boxplus R(I \otimes T_4),$$

and hence

$$X = T_3 Z_1 B_1^{o\prime} + A_2' Z_2 B_2^{o\prime} + T_4 Z_3.$$

In § 5 this result is extended. Furthermore, if, in addition to $R(B_1) \subseteq R(B_2)$, the condition $R(A_2') \subseteq R(A_1')$ exists, the solution to (2.2) and (2.3) reduces to

$$X = (A_2' : (A_1')^o)^o Z_1 B_1^{o\prime} + A_2' Z_2 B_2^{o\prime} + (A_1')^o Z_3. \tag{2.4}$$

3. The equation $A_1X_1B_1 + A_2X_2B_2 = 0$.

In this section we consider the following equation in X_1 and X_2:

$$A_1 X_1 B_1 + A_2 X_2 B_2 = 0, \tag{3.1}$$

where A_i: $p \times q_i$, B_i: $k_i \times n$, and $i = 1, 2$. A solution to (3.1) was first derived by Baksalary and Kala [1], and here we give an alternative solution that is based on the results in the preceding section.

THEOREM 2. *A representation of the general solution to* (3.1) *is*

$$X_1 = -A_1^- A_2(A_2'A_1^o : (A_2')^o)^o Z_3(B_2(B_1')^o)^{o\prime} B_2 B_1^- + (A_1')^o Z_1 + A_1' Z_2 B_1^{o\prime},$$

$$X_2 = (A_2'A_1^o : (A_2')^o)^o Z_3(B_2(B_1')^o)^{o\prime} + A_2'A_1^o Z_4 B_2^{o\prime} + (A_2')^o Z_5$$

or

$$X_1 = -A_1^- A_2(A_2'A_1^o)^o(A_2'A_1^o)^{o\prime} A_2' Z_6(B_2(B_1')^o)^{o\prime} B_2 B_1^- + (A_1')^o Z_1 + A_1' Z_2 B_1^{o\prime},$$

$$X_2 = (A_2'A_1^o)^o((A_2'A_1^o)^{o\prime} A_2')^o Z_5 + (A_2'A_1^o)^o(A_2'A_1^o)^{o\prime} A_2' Z_6(B_2(B_1')^o)^{o\prime} + A_2'A_1^o Z_4 B_2^{o\prime},$$

where the Z's are arbitrary matrices.

Proof. We are going to show that to an inhomogeneous equation in X, $AXB = C$, there exists a solution if and only if $R(C) \subseteq R(A)$ and $R(C') \subseteq R(B')$. Moreover, by aid of the particular solution $A^- CB^-$ and Theorem 1(i), a representation of the general solution is given by

$$X = A^- CB^- + (A')^o Z_1 + A' Z_2 B^{o\prime}.$$

From the above it follows that (3.1) is equivalent to

$$A_1^{o\prime} A_2 X_2 B_2 = 0, \tag{3.2}$$

$$A_2 X_2 B_2(B_1')^o = 0, \tag{3.3}$$

$$X_1 = -A_1^- A_2 X_2 B_2 B_1^- + (A_1')^o Z_1 + A_1' Z_2 B_1^{o\prime}. \tag{3.4}$$

Equations (3.2) and (3.3) do not depend on X_1 and (2.4) implies the representation of X_2, which is inserted in (3.4).

The alternative representation is readily established by using Theorem 1(i) twice. When solving (3.2),

$$X_2 = (A_2'A_1^o)^o Z_3 + A_2'A_1^o Z_4 B_2^{o\prime} \tag{3.5}$$

is obtained, and inserting (3.5) in (3.3) leads to the equation

$$A_2(A_2'A_1^o)^o Z_3 B_2(B_1')^o = 0.$$

Hence

$$Z_3 = ((A_2'A_1^o)^{o\prime} A_2')^o Z_5 + (A_2'A_1^o)^{o\prime} A_2' Z_6(B_2(B_1')^o)^{o\prime}, \tag{3.6}$$

which, in turn, is inserted into (3.5). X_1 follows once again from (3.4). □

Remark. Note that $A_2(A_2'A_1^o)^o$ in (3.6) generates $R(A_1) \cap R(A_2)$. In the complex case, in (3.2) and (3.3) we must use $\bar{A}_1^{o\prime}$ and $(\bar{B}_1')^o$ instead of $A_1^{o\prime}$ and $(B_1')^o$, respectively, and in (3.4), $(\bar{A}_1')^o$ and $(\bar{B}_1)^{o\prime}$ instead of $(A_1')^o$ and $B_1^{o\prime}$, respectively. For the alternative representations some additional alterations are needed, which follow from the remark after Theorem 1.

Parenthetically, we may note that the ideas for obtaining the alternative representation for X_2 can also be applied to (2.2) and (2.3), since from Theorem 1(i) it follows that

$$X = (A_1')^o Z_1 + A_1' Z_2 B_1^{o\prime}, \tag{3.7}$$

and plugging this into (2.3) implies that

$$A_2(A_1')^O Z_1 B_2 + A_2 A_1' Z_2 B_1^{O\prime} B_2 = 0, \tag{3.8}$$

which in turn determines Z_1 and Z_2.

As with Theorem 2, the next lemma, which will be used in § 4, can be verified. The modifications for treating complex matrices are also straightforward.

LEMMA. *Let A_i: $p \times q_i$, B_i: $k_i \times n$, and $i = 1, 2, 3, 4$. Also let the Z's represent arbitrary matrices, and let*

$$A_1 X_1 B_1 + A_2 X_2 B_2 + A_3 X_3 B_3 + A_4 X_4 B_4 = 0$$

be an equation in $X_1 - X_4$, where $R(B_4') \subseteq R(B_3') \subseteq R(B_1')$ and $R(B_4') \subseteq R(B_2') \subseteq R(B_1')$. Then

$$X_1 = -A_1^- A_2 X_2 B_2 B_1^- - A_1^- A_3 X_3 B_3 B_1^- - A_1^- A_4 X_4 B_4 B_1^- + (A_1')^O Z_1 + A_1' Z_2 B_1^{O\prime},$$

$$X_2 = -C_1^- C_2 X_3 B_3 B_2^- - C_1^- C_3 X_4 B_4 B_2^- + (C_1')^O Z_3 + C_1' Z_4 B_2^{O\prime},$$

$$X_3 = (C_2')^O Z_5 + C_2' Y (B_3 (B_2')^O)^{O\prime},$$

$$X_4 = -D_2^- D_1 Y E_1 B_4^- + (D_2')^O Z_6 + D_2' Z_7 B_4^{O\prime},$$

$$Y = (D_1' D_2^O : (D_1')^O)^O Z_8 (E_1 (B_4')^O)^{O\prime} + D_1' D_2^O Z_9 E_1^{O\prime} + (D_1')^O Z_{10},$$

where

$$C_1 = A_1^{O\prime} A_2, \quad C_2 = A_1^{O\prime} A_3, \quad C_3 = A_1^{O\prime} A_4,$$

$$D_1 = C_1^{O\prime} C_2 C_2', \quad D_2 = C_1^{O\prime} C_3, \quad E_1 = (B_3 (B_2')^O)^{O\prime}. \qquad \square$$

4. The equations $A_1XB_1 = 0$, $A_2XB_2 = 0$, $A_3XB_3 = 0$. Here we discuss the following equations in X:

$$A_1 X B_1 = 0, \tag{4.1}$$

$$A_2 X B_2 = 0, \tag{4.2}$$

$$A_3 X B_3 = 0, \tag{4.3}$$

where A_i: $p_i \times q$ and B_i: $k \times n_i$, $i = 1, 2, 3$. This is a much more complicated equation system than (2.2) and (2.3). For (2.2) and (2.3), a general representation of X was given as a sum $\Sigma_i T_i Z_i S_i$, which arose from a decomposition of tensor spaces. However, for (4.1)–(4.3), we cannot write X as a sum with arbitrary Z matrices. Instead, some restrictions on these matrices must be imposed. The following example illustrates this.

Consider the equation in X: 2×2 given by

$$(0:\ 1) X (1:\ -1)' = 0, \tag{4.4}$$

$$(1:\ 1) X (0:\ 1)' = 0, \tag{4.5}$$

$$(1:\ 2) X (1:\ 1)' = 0, \tag{4.6}$$

which, by using the Kronecker product, can be written

$$\begin{pmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \\ 1 & 2 & 1 & 2 \end{pmatrix} \operatorname{vec}(X) = 0.$$

The rank of the matrix describing the constraints equals 3. Hence, a general solution is

$$\operatorname{vec}(X) = (-3{:}\ 1{:}\ -1{:}\ 1)'Z,$$

and this solution is impossible to write as $X = SZT'$ for some vectors S and T. On the other hand, it is still possible to find a matrix satisfying the above equations, and

$$X = (-1{:}\ 1)'Z(1{:}\ 1) + (1{:}\ 0)'Z(-2{:}\ 0)$$

is one choice of a solution. Thus the technique of solving linear equations in § 2 is not sufficient for this section. However, it is still fruitful when combining it with the technique applied in § 3.

Applying Theorem 1 (ii) to (4.4) and (4.5), we get a solution which can be inserted in (4.6) and then the following equation in $Z_1 - Z_4$ is obtained:

$$(4.7)\qquad A_3T_1Z_1B_2^{O\prime}B_3 + A_3T_2Z_2(B_1{:}\ B_2)^{O\prime}B_3 + A_3T_3Z_3B_1^{O\prime}B_3 + A_3T_4Z_4B_3 = 0,$$

where $T_1 - T_4$ are as defined in Theorem 1(ii). Since $R(B_3(B_1 : B_2)^O) \subseteq R(B_3B_1^O) \subseteq R(B_3')$ and $R(B_3'(B_1 : B_2)^O) \subseteq R(B_3'B_2^O) \subseteq R(B_3')$, a solution to (4.4) is found with the help of the lemma in § 3, and thus the next theorem is established.

THEOREM 3. *A representation of a general solution to* (4.1)–(4.3) *is given by*

$$X = T_1Z_1B_2^{O\prime} + T_2Z_2(B_1 : B_2)^{O\prime} + T_3Z_3B_1^{O\prime} + T_4Z_4,$$

where $T_1 - T_4$ *are defined in Theorem* 1(ii) *and* $Z_1 - Z_4$ *are obtained by solving* (4.7). □

5. A hierarchical structure. We are going to prove the following theorem, which, among others, extends (2.4).

THEOREM 4. *A representation of the solution to*

$$A_iXB_i = 0, \qquad i = 1, 2, \dots, s,$$

when the nested subspace condition $R(B_s) \subseteq R(B_{s-1}) \subseteq \cdots \subseteq R(B_1)$ *holds, is given by* (*set* $H_i = A_1'{:}\ A_2'{:}\ \cdots : A_i'$)

$$X = H_s^OZ_1B_s' + \sum_{i=1}^{s-1} H_i^OZ_{i+1}(B_i^O : B_{i+1})^{O\prime} + Z_{s+1}B_1^{O\prime}$$

or

$$X = H_s^OZ_1 + \sum_{i=2}^{s} (H_{i-1} : H_i^O)^OZ_iB_i^{O\prime} + Z_{s+1}B_1^{O\prime},$$

where the Z's are arbitrary matrices.

Proof. Before verifying the theorem, we note that the condition $R(B_s) \subseteq R(B_{s-1}) \subseteq \cdots \subseteq R(B_1)$ enables the solution to be written as a sum $\sum_i T_iZ_iS_i$, which implies that the solution has an interpretable structure. In fact, in order to write down a relatively nice solution, we can relax the condition and just assume that the subspaces commute, i.e., $R(B_i) = R(B_i) \cap R(B_j) \boxplus R(B_i) \cap O(B_j)$ for all i, j. However, the inclusion criterion seems to have more natural applications (for example, for multivariate linear models in statistics, see von Rosen [10]) and therefore we present a solution based on this criterion.

In order to verify the theorem, for notational convenience, we use A_i', B_i, and H_i, respectively, instead of $R(A_i')$, $R(B_i)$, and $R(H_i)$. Moreover, the tensor product is denoted $\otimes$ and, as an example, the tensor product of B_i and A_i' is denoted $B_i \otimes A_i'$. Let

V_1 and V_2 represent the whole space to which B_i and A'_i, respectively, belong. By rewriting $A_i X B_i$ into a vectorized form, it follows that the theorem is true if it can be shown that ($^\perp$ denotes the orthogonal complement)

$$\begin{aligned}(5.1)\qquad \left(\sum_{i=1}^{s} B_i \otimes A'_i\right) &= B_s \otimes H_s^\perp + \boxplus_{i=1}^{s-1} (B_i \cap B_{i+1}^\perp \otimes H_i^\perp) \boxplus B_1^\perp \otimes V_2 \\ &= V_1 \otimes H_s^\perp + \boxplus_{i=2}^{s} (B_i^\perp \otimes H_i \cap H_{i-1}^\perp) \boxplus B_1^\perp \otimes A'_1.\end{aligned}$$

By applying the orthomodular law (Birkhoff [4]), we get

$$B_i = B_s + \boxplus_{i=1}^{s-1} B_i \cap B_{i+1}^\perp, \qquad i = 1, 2, \ldots, s-1,$$

and thus

$$\begin{aligned}(5.2)\qquad \sum_{i=1}^{s} B_i \otimes A'_i &= B_s \otimes H_s + \boxplus_{i=1}^{s-1} \boxplus_{j=1}^{s-1} B_j \cap B_{j+1}^\perp \otimes A'_i \\ &= B_s \otimes H_s + \boxplus_{i=1}^{s-1} B_i \cap B_{i+1}^\perp \otimes H_i,\end{aligned}$$

which obviously is orthogonal to the first statement in (5.1). Moreover, summing the first statement in (5.1) with (5.2) gives us the whole space. Hence an orthogonal complement to $\sum B_i \otimes A'_i$ has been found. The second statement in the theorem follows by straightforward manipulations and by noting that

$$H_{i-1}^\perp = H_i^\perp \boxplus H_i \cap H_{i-1}^\perp. \qquad \square$$

REFERENCES

[1] J. K. Baksalary and R. Kala, *The matrix equation $AXB + CYD = E$*, Linear Algebra Appl., 30 (1980), pp. 141–147.

[2] L. Banken, *On the reduction of the general MANOVA model*, Tech. Rep., Univ. of Trier, Trier, Germany, 1984.

[3] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, John Wiley, New York, 1974.

[4] G. Birkhoff, *Lattice Theory*, 3rd ed., American Mathematical Society, Providence, RI, 1967.

[5] S. K. Mitra, *Common solutions to a pair of linear matrix equations $A_1XB_1 = C_1$ and $A_2XB_2 = C_2$*, Proc. Cambridge Philos. Soc., 74 (1973), pp. 213–216.

[6] S. K. Mitra, *A pair of simultaneous linear matrix equations $A_1XB_1 = C_1$ and $A_2XB_2 = C_2$ and a matrix programming problem*, Linear Algebra Appl., 131 (1990), pp. 107–123.

[7] K. Nordström and D. von Rosen, *Algebra of subspaces with applications to problems in statistics*, in Proc. Second Internat. Tampere Conference in Statistics, T. Pukkila and S. Puntanen, eds., University of Tampere, Tampere, Finland, 1987, pp. 603–614.

[8] R. F. Potthoff and S. N. Roy, *A generalized multivariate analysis of variance model useful especially for growth curve problems*, Biometrika, 51 (1964), pp. 313–326.

[9] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York, 1971.

[10] D. von Rosen, *Maximum likelihood estimators in multivariate linear normal models*, J. Multivariate Anal., 31 (1989), pp. 187–200.

[11] A. P. Verbyla and W. N. Venables, *An extension of the growth curve model*, Biometrika, 75 (1988), pp. 125–138.

SPECTRAL PROPERTIES OF PRECONDITIONED RATIONAL TOEPLITZ MATRICES*

TAKANG KU† AND C.-C. JAY KUO†

Abstract. Various Toeplitz preconditioners P_N have recently been proposed so that an $N \times N$ symmetric positive definite Toeplitz system T_N $\mathbf{x} = \mathbf{b}$ can be solved effectively by the preconditioned conjugate gradient (PCG) method. It has been proven that if T_N is generated by a positive function in the Wiener class, the eigenvalues of the preconditioned matrices $P_N^{-1}T_N$ are clustered between $(1-\epsilon, 1+\epsilon)$ except for a fixed number independent of N. In this research, the spectra of $P_N^{-1}T_N$ are characterized more precisely for rational Toeplitz matrices T_N with preconditioners proposed by Strang [*Stud. Appl. Math.*, 74 (1986), pp. 171–176] and Ku and Kuo [*IEEE Trans. Signal Process.*, 40 (1992), pp. 129–141]. The eigenvalues of $P_N^{-1}T_N$ are classified into two classes, i.e., the outliers and the clustered eigenvalues, depending on whether they converge to 1 asymptotically. It is proved that the number of outliers depends on the order of the rational generating function, and the clustering radius ϵ is proportional to the magnitude of the last element in the generating sequence used to construct these preconditioners. For the special case with T_N generated by a geometric sequence, this approach can be used to determine the exact eigenvalue distribution of $P_N^{-1}T_N$ analytically.

Key words. Toeplitz matrix, preconditioned conjugate gradient method, rational generating function

AMS(MOS) subject classifications. 65F10, 65F15

1. Introduction.

The system of linear equations associated with a symmetric positive definite (SPD) Toeplitz matrix arises in many applications, such as time series analysis and digital signal processing. The $N \times N$ symmetric Toeplitz system $T_N\mathbf{x} = \mathbf{b}$ is conventionally solved by algorithms based on the Levinson recursion formula [10], [16] with $O(N^2)$ operations. Superfast algorithms with $O(N \log^2 N)$ complexity have been studied intensively in the last ten years [1], [2], [3], [13]. More recently, Strang [19] proposed using an iterative method, i.e., the preconditioned conjugate gradient (PCG) method, to solve SPD Toeplitz systems and, as a consequence, the design of effective Toeplitz preconditioners has received much attention.

Strang's preconditioner S_N [19] is obtained by preserving the central half-diagonals of T_N and using them to form a circulant matrix. Since S_N is circulant, the matrix-vector product $S_N^{-1}\mathbf{v}$ can be conveniently computed via fast Fourier transform (FFT) with $O(N \log N)$ operations. It has been shown by R. Chan and Strang [5], [7] that if T_N is generated by a positive function in the Wiener class, the eigenvalues of the preconditioned matrices $P_N^{-1}T_N$ are clustered between $(1-\epsilon, 1+\epsilon)$ except for a fixed number independent of N. Another preconditioner C_N was proposed by T. Chan [8] and is defined to be the circulant matrix that minimizes the Frobenius norm $||R_N - T_N||_F$ over all circulant matrices R_N of size $N \times N$. This turns out to be a simple optimization problem, and the elements of C_N can be computed directly from the elements of T_N. The spectrum of $C_N^{-1}T_N$ is asymptotically equivalent to that of $S_N^{-1}T_N$ [6], and thus C_N and S_N have similar asymptotic behavior. In addition to preconditioners in circulant matrix form, preconditioners in skew-circulant matrix form [9] have been studied by Huckle [14]. We recently proposed a general approach

* Received by the editors September 11, 1990; accepted for publication (in revised form) March 3, 1991. This work was supported by the University of Southern California Faculty Research and Innovation Fund and by a National Science Foundation Research Initiation Award (ASC-9009323).

† Signal and Image Processing Institute and Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, California 90089-2564 (tkku@sipi.usc.edu and cckuo@sipi.usc.edu).

for constructing Toeplitz preconditioners [15]. Under this framework, preconditioners in circulant and skew-circulant matrix forms can be viewed as special cases and, more interestingly, preconditioners that are neither circulant nor skew-circulant can also be derived.

In [15], four new preconditioners $K_{i,N}$, $i = 1, 2, 3, 4$ were constructed, and it was demonstrated numerically that they have better convergence performances than other preconditioners for rational Toeplitz matrices. It was also observed in [15] that for T_N generated by a positive rational function of order (p, q) in the Wiener class, the spectra of the preconditioned matrix $P_N^{-1}T_N$ with preconditioners S_N and $K_{i,N}$, $i = 1, 2, 3, 4$, have strong regularities. These regularities are stated as follows. Let the eigenvalues of $P_N^{-1}T_N$ be classified into two classes, i.e., the outliers and the clustered eigenvalues, depending on whether they converge to 1 asymptotically. Then, (1) the number of outliers is at most $2\max(p, q)$; and (2) the clustered eigenvalues are confined in an interval $(1 - \epsilon, 1 + \epsilon)$ with the radius ϵ proportional to the magnitude of the last element in the generating sequence used to construct the preconditioner. The main objective of this research is to prove these two spectral properties analytically.

With the above spectral regularities, the number of iterations required to reduce the norm of the residual $||\mathbf{b} - T_N\mathbf{x}_k||$ by a constant factor does not increase with the problem size N so that the solution of the system $T_N\mathbf{x} = \mathbf{b}$ can be accomplished with $\max(p, q) \times O(N \log N)$ operations. In addition, the superior performance of preconditioners $K_{i,N}$ can be easily explained by these spectral regularities. That is, for T_N generated by a positive rational function in the Wiener class, the last elements used to construct $K_{i,N}$ and S_N are, respectively, t_N and $t_{\lceil N/2\rceil}$ so that the corresponding radii are $\epsilon_K = O(|t_N|)$ and $\epsilon_S = O(|t_{\lceil N/2\rceil}|)$. Since $O(|t_N|) << O(|t_{\lceil N/2\rceil}|)$ for sufficiently large N, the PCG method with preconditioners $K_{i,N}$ converges faster than with preconditioner S_N.

We should point out that the first spectral property was recently proved by Trefethen. In [23], he used the theory of CF (Carathéodory and Fejér) approximation [22] to show that $S_N^{-1}T_N$ has at most $1 + 2\max(p, q)$ distinct eigenvalues asymptotically. A different approach is adopted in this paper to prove this property for both $S_N^{-1}T_N$ and $K_{i,N}^{-1}T_N$ (see Lemmas 2 and 8). Besides, since the first property only characterizes the spectrum of $P_N^{-1}T_N$ for infinite N, whereas the second property characterizes the spectrum of $P_N^{-1}T_N$ for both finite and infinite N, our results have a greater generality.

There exist direct methods that solve rational Toeplitz systems with $\max(p, q) \times O(N)$ operations [11], [24], [25]. However, the PCG method has three advantages compared with these direct methods. First, to implement the PCG algorithm, we only need a finite segment of the generating sequence t_n, $n = 0, 1, \ldots, N - 1$, which is provided by the problem, rather than the precise formula of the rational generating function. Second, the PCG method can be easily parallelized due to the parallelism provided by FFT, and it is possible to reduce the time complexity to $\max(p, q) \times O(\log N)$. In contrast, these direct methods are sequential algorithms, and the time complexity cannot be further reduced. Third, the PCG method is more widely applicable. For example, it can also be applied to Toeplitz systems with nonrational Toeplitz generating functions or those arising from the multidimensional space.

This paper is organized as follows. In §2, we briefly review the construction of preconditioners $K_{i,N}$ and summarize some of their spectral properties studied in [15]. In §§3 and 4, we prove the desired spectral properties of $K_{i,N}^{-1}T_N$ described above. The main idea is to transform the original generalized eigenvalue problem to an equivalent problem with nearly banded Toeplitz matrices. A similar approach is used to study the spectral properties of $S_N^{-1}T_N$, which is presented in §5. In §6, we use the analysis in §§3–5 to determine the analytical eigenvalue distributions of $K_{i,N}^{-1}T_N$ and $S_N^{-1}T_N$ for Toeplitz matrices with a geometric generating sequence.

2. Construction and spectral properties of Toeplitz preconditioners $K_{i,N}, i = 1, 2, 3, 4$. Let T_m be a sequence of $m \times m$ symmetric positive definite Toeplitz matrices with generating sequence t_n. Then,

$$T_N = \begin{bmatrix} t_0 & t_1 & \cdot & t_{N-2} & t_{N-1} \\ t_1 & t_0 & t_1 & \cdot & t_{N-2} \\ \cdot & t_1 & t_0 & \cdot & \cdot \\ t_{N-2} & \cdot & \cdot & \cdot & t_1 \\ t_{N-1} & t_{N-2} & \cdot & t_1 & t_0 \end{bmatrix}.$$

Preconditioners $K_{i,N}$, $i = 1, 2, 3, 4$, for T_N are constructed by relating T_N to a $2N \times 2N$ circulant matrix R_{2N},

$$R_{2N} = \begin{bmatrix} T_N & \triangle T_N \\ \triangle T_N & T_N \end{bmatrix},$$

where $\triangle T_N$ is determined by the elements of T_N to make R_{2N} circulant, i.e.,

$$\triangle T_N = \begin{bmatrix} c & t_{N-1} & \cdot & t_2 & t_1 \\ t_{N-1} & c & t_{N-1} & \cdot & t_2 \\ \cdot & t_{N-1} & c & \cdot & \cdot \\ t_2 & \cdot & \cdot & \cdot & t_{N-1} \\ t_1 & t_2 & \cdot & t_{N-1} & c \end{bmatrix}, \tag{2.1}$$

with a constant c. If the behavior of the sequence t_n is known, we choose c to be t_N. Otherwise, $c = 0$.

Consider the following augmented circulant system:

$$\begin{bmatrix} T_N & \triangle T_N \\ \triangle T_N & T_N \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{b} \end{bmatrix}. \tag{2.2}$$

The solution of the above circulant system can be computed efficiently via FFT with $O(N \log N)$ operations. Since (2.2) is equivalent to

$$(T_N + \triangle T_N)\mathbf{x} = \mathbf{b},$$

this implies that $(T_N + \triangle T_N)^{-1}\mathbf{b}$ can be computed efficiently and that

$$K_{1,N} = T_N + \triangle T_N$$

can be used as a preconditioner for T_N. Three other preconditioners can be constructed in a similar way by assuming negative, even and odd periodicities for $\mathbf{x}$ and $\mathbf{b}$. We summarize the augmented systems and the corresponding preconditioners as follows:

$$\begin{bmatrix} T_N & \triangle T_N \\ \triangle T_N & T_N \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ -\mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ -\mathbf{b} \end{bmatrix} \quad \text{and} \quad K_{2,N} = T_N - \triangle T_N,$$

$$\begin{bmatrix} T_N & \triangle T_N \\ \triangle T_N & T_N \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ J_N\mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ J_N\mathbf{b} \end{bmatrix} \quad \text{and} \quad K_{3,N} = T_N + J_N\triangle T_N,$$

$$\begin{bmatrix} T_N & \triangle T_N \\ \triangle T_N & T_N \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ -J_N\mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ -J_N\mathbf{b} \end{bmatrix} \quad \text{and} \quad K_{4,N} = T_N - J_N\triangle T_N,$$

where J_N is the $N \times N$ symmetric elementary matrix which has, by definition, ones along the secondary diagonal and zeros elsewhere ($J_{N,i,j} = 1$ if $i + j = N + 1$ and $J_{N,i,j} = 0$ if $i + j \neq N + 1$).

Since preconditioners $K_{i,N}$, $i = 1, 2, 3, 4$, correspond to $2N$-circulant systems, the matrix-vector product $K_{i,N}^{-1}\mathbf{v}$ for an arbitrary $\mathbf{v}$ can be achieved via $2N$-point FFT with $O(N \log N)$ operations. However, we should point out that $K_{1,N}$ is circulant and $K_{2,N}$ is skew-circulant so that $K_{1,N}^{-1}\mathbf{v}$ and $K_{2,N}^{-1}\mathbf{v}$ can be computed via N-point FFT. Although preconditioners $K_{3,N}$ and $K_{4,N}$ are neither circulant nor skew-circulant, $K_{3,N}^{-1}\mathbf{v}$ and $K_{4,N}^{-1}\mathbf{v}$ can be computed via N-point fast cosine and sine transforms, respectively. The operation count for N-point fast cosine (or sine) transform is approximately equal to that of N-point FFT in both the order and the proportional constant [17], [18], [27]. Therefore, the computational cost for the preconditioning step $K_{i,N}^{-1}\mathbf{v}$ with $i = 1, 2, 3, 4$ is about the same. For more details in implementing the PCG algorithm, we refer to [15].

To understand the relationship between the spectra of $K_{i,N}^{-1}T_N$, $i = 1, 2, 3, 4$, we rewrite the eigenvalues of $K_{i,N}^{-1}T_N$ as

$$\begin{aligned} [\lambda(K_{i,N}^{-1}T_N)]^{-1} &= \lambda(T_N^{-1}(T_N + K_{i,N} - T_N)) = \lambda(I + T_N^{-1}(K_{i,N} - T_N)) \\ &= 1 + \lambda(T_N^{-1}(K_{i,N} - T_N)), \end{aligned} \tag{2.3}$$

and examine the relationship between the spectra of $T_N^{-1}(K_{i,N} - T_N)$. This is characterized by the following theorem.

THEOREM 1. *Let Q_i be the set of the absolute values of the eigenvalues of $T_N^{-1}(K_{i,N} - T_N)$, i.e.,*

$$Q_i = \{|\lambda| : (K_{i,N} - T_N)\mathbf{x} = \lambda T_N \mathbf{x}\}, \qquad i = 1, 2, 3, 4.$$

Then, $Q_1 = Q_2 = Q_3 = Q_4$.

Proof. See [15] for the proof. □

The above theorem can be stated alternatively as follows. Let λ be an arbitrary eigenvalue of $T_N^{-1}(K_{i,N} - T_N)$; then there exists an eigenvalue of $T_N^{-1}(K_{j,N} - T_N)$, where $j \neq i$, with magnitude $|\lambda|$. From (2.3), spectra of $T_N^{-1}(K_{i,N} - T_N)$ clustered around zero are equivalent to those of $K_{i,N}^{-1}T_N$ clustered around unity. Since spectra of $T_N^{-1}(K_{i,N} - T_N)$ are clustered in a very similar pattern, so are those of $K_{i,N}^{-1}T_N$.

We assume that the generating sequence t_n for the sequence of Toeplitz matrices T_m satisfies the following two conditions:

$$\sum_{-\infty}^{\infty} |t_n| < \infty, \tag{2.4}$$

$$T(e^{i\theta}) = \sum_{-\infty}^{\infty} t_n e^{-in\theta} \geq \delta > 0 \quad \forall\theta, \tag{2.5}$$

and the resulting matrices are said to be generated by a positive function in the Wiener class. Since $T(e^{i\theta})$ describes the asymptotic eigenvalue distribution of T_m, the above conditions assume that the eigenvalues of T_m are bounded and uniformly positive, asymptotically. With (2.4) and (2.5), two spectral properties of $K_{i,N}^{-1}T_N$ are derived.

THEOREM 2. *Preconditioners $K_{i,N}$, $i = 1,2,3,4$, for symmetric positive definite Toeplitz matrices T_N with the generating sequence satisfying* (2.4) *and* (2.5) *are uniformly positive definite and bounded for sufficiently large N.*

Proof. See [15] for the proof. □

THEOREM 3. *Let T_N be the $N \times N$ matrix in a sequence of $m \times m$ symmetric positive definite Toeplitz matrices T_m with the generating sequence satisfying* (2.4) *and* (2.5). *The eigenvalues of the matrix $T_N^{-1}(K_{i,N} - T_N)$ are clustered between $(-\epsilon, +\epsilon)$ except for a finite number of outliers for sufficiently large $N(\epsilon)$.*

Proof. See [15] for the proof. □

Theorems 2 and 3 hold for both rational and nonrational Toeplitz matrices satisfying (2.4) and (2.5). However, when T_N is additionally rational, we are able to obtain stronger results and characterize the spectra of $K_{i,N}^{-1}T_N$ more precisely. In §§3 and 4, we focus on the spectrum of $K_{1,N}^{-1}T_N$, from which the spectra of $K_{i,N}^{-1}T_N$, $i = 2,3,4$, can be estimated based on Theorem 1.

3. Rational generating functions for $\triangle T_N$. Due to (2.3), the spectral properties of $K_{1,N}^{-1}T_N$ can be determined by examining those of $T_N^{-1}\triangle T_N$, where $\triangle T_N$ is given in (2.1) with $c = t_N$. Let t_n, $-\infty < n < \infty$, be the generating sequence of a sequence of $m \times m$ Toeplitz matrices T_m. The Laurent series

$$T(z) = \sum_{n=-\infty}^{\infty} t_n z^{-n}$$

is known as the generating function of these matrices. If matrices T_m are symmetric, we decompose $T(z)$ into

$$T(z) = T_+(z^{-1}) + T_+(z), \tag{3.1}$$

where

$$T_+(z^{-1}) = \frac{t_0}{2} + \sum_{n=1}^{\infty} t_n z^{-n}. \tag{3.2}$$

Thus $T(z)$ is completely characterized by $T_+(z^{-1})$. Additionally, if

$$T_+(z^{-1}) = \frac{A(z^{-1})}{B(z^{-1})} = \frac{a_0 + a_1 z^{-1} + \cdots + a_p z^{-p}}{b_0 + b_1 z^{-1} + \cdots + b_q z^{-q}}, \tag{3.3}$$

where $b_0 = 1$, $a_p b_q \neq 0$, and polynomials $A(z^{-1})$ and $B(z^{-1})$ have no common factor, we call T_m the rational Toeplitz matrices generated by a rational function of order (p,q). From (3.1) and (3.3), we have

$$T(z) = \frac{A(z^{-1})}{B(z^{-1})} + \frac{A(z)}{B(z)}. \tag{3.4}$$

It is well known [12] that there exists an isomorphism between the ring of the power series $P(z^{-1}) = \sum_{n=0}^{\infty} p_n z^{-n}$ (or $P(z)$) and the ring of the semi-infinite lower (or upper) triangular Toeplitz matrices with $p_0, p_1, \cdots, p_n, \cdots$ as the first column (or row). The power series multiplication is isomorphic to matrix multiplication. By applying the isomorphism to (3.4) and focusing on the leading $N \times N$ blocks of the corresponding matrices, we derive the following relationship [12]:

$$T_N = L_a L_b^{-1} + U_a U_b^{-1}, \tag{3.5}$$

where L_a (or U_a) is an $N \times N$ lower (or upper) triangular Toeplitz matrix with first N coefficients in $A(z^{-1})$ as its first column (or row). Matrices L_b and U_b are defined similarly with respect to $B(z^{-1})$. We can also establish an expression similar to (3.5) for $\triangle T_N$. To do so, we first note that the sequence t_n is recursively defined for large n. This is stated as follows.

LEMMA 1. *The sequence t_n generated by* (3.2) *and* (3.3) *follows the recursion,*

$$t_{n+1} = -(b_1 t_n + b_2 t_{n-1} + \cdots + b_q t_{n-q+1}), \qquad n \geq \max(p,q). \tag{3.6}$$

Proof. From (3.2) and (3.3), we have

$$\left(\frac{t_0}{2} + \sum_{n=1}^{\infty} t_n z^{-n}\right)(b_0 + b_1 z^{-1} + \cdots + b_q z^{-q}) = a_0 + a_1 z^{-1} + \cdots + a_p z^{-p}.$$

The proof is completed by comparing the coefficients of the above equation. □

With Lemma 1, the number of outliers of $T_N^{-1} \triangle T_N$ is determined by the following lemma.

LEMMA 2. *Let T_N be an $N \times N$ symmetric Toeplitz matrix generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *and the corresponding generating sequence satisfies* (2.4) *and* (2.5). *$T_N^{-1} \triangle T_N$ has asymptotically at most $2\max(p,q)$ nonzero eigenvalues* (*outliers*).

Proof. Let us define a matrix

$$\triangle E_N = \triangle F_N + \triangle F_N^T,$$

where

$$\triangle F_N = \begin{bmatrix} t_N & t_{N-1} & \cdot & t_2 & t_1 \\ t_{N+1} & t_N & t_{N-1} & \cdot & t_2 \\ \cdot & t_{N+1} & t_N & \cdot & \cdot \\ t_{2N-2} & \cdot & \cdot & \cdot & t_{N-1} \\ t_{2N-1} & t_{2N-2} & \cdot & t_{N+1} & t_N \end{bmatrix}.$$

Since elements t_n in $\triangle F_N$ satisfy (3.6), there are at most $\max(p,q)$ independent rows in $\triangle F_N$ and therefore, the rank of $\triangle E_N$ is at most $2\max(p,q)$.

Let $\triangle P_N = \triangle E_N - \triangle T_N$; it is easy to verify that the l_1 and l_∞ norms of $\triangle P_N$ are both less than

$$\tau_K = 2 \sum_{n=N}^{2N-1} |t_n|.$$

Consequently, we have

$$||\triangle P_N||_2 \leq (||\triangle P_N||_1 ||\triangle P_N||_\infty)^{1/2} \leq \tau_K.$$

Since τ_K goes to zero as N goes to infinity due to (2.4), and since the eigenvalues of T_N^{-1} are bounded due to (2.5), the spectra of $T_N^{-1} \triangle T_N$ and $T_N^{-1} \triangle E_N$ are asymptotically equivalent. It follows that both $T_N^{-1} \triangle E_N$ and $T_N^{-1} \triangle T_N$ have at most $2\max(p,q)$ nonzero eigenvalues asymptotically. □

As a consequence of Lemma 2, $T_N^{-1} \triangle T_N$ has at least $N - 2\max(p,q)$ eigenvalues converging to zero as the problem size N becomes large. For the rest of this section and in §4, we study the clustering property of these eigenvalues. Our approach is outlined as follows. First, we associate $\triangle T_N$ with some appropriate rational generating function $\tilde{T}(z) = \tilde{T}_+(z^{-1}) + \tilde{T}_+(z)$. The forms of $\tilde{T}_+(z^{-1})$ for $p \leq q$ and $p > q$ are given in Lemmas 3 and 4, respectively. We then transform the generalized eigenvalue problem involving $T_N^{-1}\triangle T_N$ into another generalized eigenvalue problem involving $Q_N^{-1}\triangle Q_N$. We show that Q_N and $\triangle Q_N$ are nearly banded Toeplitz matrices in Lemma 5 and examine the spectral property of $Q_N^{-1}\triangle Q_N$ in Lemma 6.

Since T_N is a symmetric rational Toeplitz matrix, and the elements of $\triangle T_N$ are those of T_N with reverse ordering, it is not surprising that $\triangle T_N$ is also generated by a certain rational function, which is determined below. Let us use the elements t_n of a given T_N with $N > \max(p, q)$ to construct a new sequence $\tilde{t}_n$. The cases $p \leq q$ and $p > q$ are considered separately.

Case 1. $p \leq q$. We choose

$$\tilde{t}_n = \begin{cases} t_{N-n}, & 0 \leq n \leq q-1, \\ -(\sum_{k=1}^{q} b_{q-k}\tilde{t}_{n-k})/b_q, & q \leq n. \end{cases} \tag{3.7}$$

Note that elements $\tilde{t}_n$ above with $n \geq q$ are obtained based on the recursion (3.6) examined from the reverse direction.

Case 2. $p > q$. We decompose $T_+(z^{-1})$ into

$$T_+(z^{-1}) = F_+(z^{-1}) + T_{1,+}(z^{-1}), \tag{3.8}$$

where

$$F_+(z^{-1}) = f_0 + f_1 z^{-1} + \cdots + f_{p-q} z^{-(p-q)} \tag{3.8a}$$

and

$$T_{1,+}(z^{-1}) = \frac{A'(z^{-1})}{B(z^{-1})} = \frac{a'_0 + a'_1 z^{-1} + \cdots + a'_s z^{-s}}{b_0 + b_1 z^{-1} + \cdots + b_q z^{-q}}, \tag{3.8b}$$

with $s < q$. Let $t_{1,n}$ be the generating sequence of $T_{1,+}(z^{-1})$. There exists a simple relationship between the elements of generating sequences for $T_+(z^{-1})$ and $T_{1,+}(z^{-1})$, i.e.,

$$t_n = \begin{cases} t_{1,n} + f_n, & 0 \leq n \leq p-q, \\ t_{1,n}, & p-q < n. \end{cases}$$

With respect to $T_{1,+}(z^{-1})$ and $F_+(z^{-1})$, we choose the corresponding $\tilde{t}_{1,n}$ and $\tilde{t}_{2,n}$, respectively, as

$$\tilde{t}_{1,n} = \begin{cases} t_{1,N-n}, & 0 \leq n \leq q-1, \\ -(\sum_{k=1}^{q} b_{q-k}\tilde{t}_{1,n-k})/b_q, & q \leq n, \end{cases}$$

and

$$\tilde{t}_{2,n} = \begin{cases} f_{N-n}, & N-p+q \leq n \leq N, \\ 0, & \text{elsewhere}. \end{cases}$$

Finally, we define

$$\tilde{t}_n = \tilde{t}_{1,n} + \tilde{t}_{2,n}. \tag{3.9}$$

We associate the sequence $\tilde{t}_n$ given by (3.7) or (3.9) with a sequence of symmetric Toeplitz matrices $\tilde{T}_m$. It is straightforward to verify that for $N > \max(p, q)$, $\tilde{T}_N = \triangle T_N$. The generating function for matrices $\tilde{T}_m$ is

$$\tilde{T}(z) = \tilde{T}_+(z^{-1}) + \tilde{T}_+(z), \quad \text{where } \tilde{T}_+(z^{-1}) = \frac{\tilde{t}_0}{2} + \sum_{n=1}^{\infty} \tilde{t}_n z^{-n}.$$

The forms of $\tilde{T}_+(z^{-1})$ with $p \le q$ and $p > q$ are described, respectively, in Lemmas 3 and 4.

LEMMA 3. *If T_N is generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *and $p \le q < N$, then $\triangle T_N$ is generated by $\tilde{T}(z)$ with*

$$\tilde{T}_+(z^{-1}) = \frac{C(z^{-1})}{D(z^{-1})} = \frac{c_0 + c_1 z^{-1} + \cdots + c_q z^{-q}}{d_0 + d_1 z^{-1} + \cdots + d_q z^{-q}}, \tag{3.10}$$

where

$$d_i = \begin{cases} b_q^{-1} b_{q-i}, & 0 \le i \le q, \\ 0, & q < i, \end{cases} \qquad c_i = \begin{cases} \sum_{j=0}^{i} d_j \tilde{t}'_{i-j}, & 0 \le i \le q, \\ 0, & q < i, \end{cases} \tag{3.10a}$$

and where

$$\tilde{t}'_n = \begin{cases} \tilde{t}_n, & n \neq 0, \\ \frac{\tilde{t}_0}{2}, & n = 0, \end{cases} \tag{3.10b}$$

with $\tilde{t}_n$ given by (3.7).

Proof. By (3.7), the sequence $\tilde{t}_n$ satisfies the recursion

$$\tilde{t}_k = -(d_1 \tilde{t}_{k-1} + d_2 \tilde{t}_{k-2} + \cdots + d_q \tilde{t}_{k-q}), \quad \text{for } k \ge q, \tag{3.11}$$

with d_i given by (3.10a). Let us define $G_k(z^{-k})$, $k > q$, as

$$G_k(z^{-k}) = (\tilde{t}_k + d_1 \tilde{t}_{k-1} + d_2 \tilde{t}_{k-2} + \cdots + d_q \tilde{t}_{k-q}) z^{-k}.$$

It is evident from (3.11) that $G_k(z^{-k}) = 0$ for $k > q$. Therefore, we have

$$\begin{aligned}(1 + d_1 z^{-1} + \cdots + d_q z^{-q}) \tilde{T}_+(z^{-1}) &= \sum_{i=0}^{q} \left(\sum_{j=0}^{i} d_j \tilde{t}'_{i-j} \right) z^{-i} + \sum_{k=q+1}^{\infty} G_k(z^{-k}) \\ &= c_0 + c_1 z^{-1} + \cdots + c_q z^{-q},\end{aligned}$$

with c_i and $\tilde{t}'_i$ defined in (3.10a) and (3.10b), respectively. This completes the proof. □

LEMMA 4. *If T_N is generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *and $q < p < N$, then $\triangle T_N$ is generated by $\tilde{T}(z)$ with*

$$\tilde{T}_+(z^{-1}) = \frac{C(z^{-1})}{D(z^{-1})} + F_+(z) z^{-N}, \tag{3.12}$$

where

$$d_i = \begin{cases} b_q^{-1} b_{q-i}, & 0 \le i \le q, \\ 0, & q < i, \end{cases} \qquad c_i = \begin{cases} \sum_{j=0}^{i} d_j \tilde{t}'_{i-j}, & 0 \le i \le q, \\ 0, & q < i, \end{cases} \tag{3.12a}$$

and where

$$(3.12b) \qquad \tilde{t}'_n = \begin{cases} \tilde{t}_n, & n \neq 0, \\ \frac{\tilde{t}_0}{2}, & n = 0, \end{cases}$$

with $\tilde{t}_n$ *given by* (3.9).

Proof. Due to (3.8), we express $\tilde{T}_+(z^{-1})$ as $\tilde{T}_+(z^{-1}) = \tilde{F}_+(z^{-1}) + \tilde{T}_{1,+}(z^{-1})$, where

$$\tilde{F}_+(z^{-1}) = \sum_{n=N-p+q}^{N} f_{N-n} z^{-n}, \qquad \tilde{T}_{1,+}(z^{-1}) = \frac{\tilde{t}_{1,0}}{2} + \sum_{n=1}^{\infty} \tilde{t}_{1,n} z^{-n}.$$

It is clear from Lemma 3 and (3.8a) that

$$\tilde{F}_+(z^{-1}) = F_+(z) z^{-N}, \qquad \tilde{T}_{1,+}(z^{-1}) = \frac{C(z^{-1})}{D(z^{-1})}.$$

Thus the proof is completed. □

We rewrite (3.12) as

$$(3.13) \quad \tilde{T}_+(z^{-1}) = \frac{C_1(z^{-1})}{D(z^{-1})}, \quad \text{with } C_1(z^{-1}) = C(z^{-1}) + D(z^{-1}) F_+(z) z^{-N}.$$

Applying the isomorphism to (3.10) or (3.13) and focusing on the leading $N \times N$ blocks of the corresponding matrices, we obtain

$$\tilde{T}_N = L_c L_d^{-1} + U_c U_d^{-1},$$

where L_c (or U_c) is an $N \times N$ lower (or upper) triangular Toeplitz matrix with the first N coefficients of $C(z^{-1})$ $(p \leq q)$ or $C_1(z^{-1})$ $(p > q)$ as its first column (or row). Matrices L_d and U_d are similarly defined with respect to $D(z^{-1})$. Since $\triangle T_N = \tilde{T}_N$, we obtain

$$(3.14) \qquad \triangle T_N = L_c L_d^{-1} + U_c U_d^{-1}.$$

4. Spectral properties of $T_N^{-1} \triangle T_N$. With the results given by (3.5) and (3.14), we then transform the generalized eigenvalue problem,

$$(4.1) \qquad \triangle T_N \mathbf{x} = \lambda T_N \mathbf{x},$$

to an equivalent generalized eigenvalue problem,

$$(4.2) \qquad \triangle Q_N \mathbf{y} = \lambda Q_N \mathbf{y},$$

where

$$(4.2a) \qquad Q_N = L_b T_N U_b = L_a U_b + L_b U_a$$

and

$$(4.2b) \qquad \triangle Q_N = L_b \triangle T_N U_b = L_b L_c L_d^{-1} U_b + L_b U_c U_d^{-1} U_b.$$

It is clear that (4.1) and (4.2) have identical eigenvalues and their eigenvectors are related via $\mathbf{x} = U_b \mathbf{y}$. The reason for (4.2) is that Q_N and $\triangle Q_N$ are nearly banded

Toeplitz matrices which can be more easily analyzed. The properties of matrices Q_N and $\triangle Q_N$ are characterized below.

LEMMA 5. *Let T_m be a sequence of $m \times m$ symmetric Toeplitz matrices generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *and the corresponding generating sequence satisfies* (2.4) *and* (2.5). *The southeast $(N - \max(p,q)) \times (N - \max(p,q))$ blocks of Q_N and $\triangle Q_N$ are symmetric banded Toeplitz matrices with generating functions*

$$Q(z) = A(z^{-1})B(z) + B(z^{-1})A(z) \tag{4.3}$$

and

$$\triangle Q(z) = b_q z^q C(z^{-1})B(z^{-1}) + b_q z^{-q} C(z)B(z), \tag{4.4}$$

respectively.

Proof. Consider two Toeplitz matrices F_N and G_N of size $N \times N$, where F_N is a lower triangular Toeplitz matrix with lower bandwidth r and the generating function $F(z^{-1})$, G_N an upper triangular Toeplitz matrix with upper bandwidth s and the generating function $G(z)$. It is easy to verify that the product $F_N G_N$, except for its northwest $r \times s$ block, is a banded Toeplitz matrix with the lower bandwidth r, upper bandwidth s, and generating function $F(z^{-1})G(z)$. We generalize the above result to $Q_N = L_a U_b + L_b U_a$ and find that the southeast $(N - \max(p,q)) \times (N - \max(p,q))$ block of Q_N is a symmetric banded Toeplitz matrix with the generating function

$$Q(z) = A(z^{-1})B(z) + B(z^{-1})A(z).$$

Since the product of lower (or upper) triangular Toeplitz matrices is commutative, we rewrite (4.2b) as

$$\triangle Q_N = \triangle Q_{1,N} + \triangle Q_{1,N}^T, \quad \text{where } \triangle Q_{1,N} = L_b L_c L_d^{-1} U_b.$$

When $p \leq q$, the product $L_b L_c L_d^{-1}$ results in a lower triangular Toeplitz matrix with the generating function $B(z^{-1})C(z^{-1})D^{-1}(z^{-1})$. The matrix $\triangle Q_{1,N}$, except for the first q columns, is a Toeplitz matrix with the generating function

$$\triangle Q_{1,N}(z^{-1}) = B(z^{-1})C(z^{-1})D^{-1}(z^{-1})B(z).$$

We use (3.10a) to relate $D(z^{-1})$ with $B(z)$, i.e.,

$$D(z^{-1}) = \sum_{n=0}^{q} d_n z^{-n} = b_q^{-1} z^{-q} \sum_{n=0}^{q} b_{q-n} z^{q-n} = b_q^{-1} z^{-q} B(z).$$

Thus $\triangle Q_{1,N}(z^{-1}) = b_q z^q B(z^{-1})C(z^{-1})$. Similarly, $\triangle Q_{1,N}^T$, except for the first q rows, is a Toeplitz matrix with the generating function $\triangle Q_{1,N}(z)$. Therefore, the southeast $(N-q) \times (N-q)$ block of $\triangle Q_N$ is a symmetric banded Toeplitz matrix with the generating function

$$\triangle Q(z) = \triangle Q_{1,N}(z^{-1}) + \triangle Q_{1,N}(z) = b_q \; (z^q B(z^{-1})C(z^{-1}) + z^{-q} B(z)C(z)),$$

where the coefficients of $C(z^{-1})$ are given in Lemma 3.

When $p > q$, the generating function of matrix L_c is $C_1(z^{-1})$ in (3.13). Consequently, $\triangle Q_{1,N}$, except for the first q columns, is a Toeplitz matrix with the generating function

$$\begin{aligned}\triangle Q_{1,N}(z^{-1}) &= B(z^{-1})C_1(z^{-1})D^{-1}(z^{-1})B(z) \\ &= B(z^{-1})C(z^{-1})D^{-1}(z^{-1})B(z) + z^{-N} B(z^{-1})F_+(z)B(z).\end{aligned}$$

Recall that the orders of polynomials $B(z)$ and $F_+(z)$ are q and $p-q$, respectively. The lowest order in z of the polynomial $z^{-N}B(z^{-1})F_+(z)B(z)$ is $-(N-p)$, and the elements of the leading $N \times N$ Toeplitz matrix generated by $z^{-N}B(z^{-1})F_+(z)B(z)$ are zeros except for the southwest p diagonals. Therefore, the matrix $\triangle Q_{1,N}$, except for the first q columns and the southwest p diagonals, is a Toeplitz matrix with the generating function

$$\triangle Q_{1,N}(z^{-1}) = B(z^{-1})C(z^{-1})D^{-1}(z^{-1})B(z).$$

Then it follows that the southeast $(N-p) \times (N-p)$ block of $\triangle Q_N$ is a symmetric banded Toeplitz matrix with the generating function

$$\triangle Q(z) = b_q \, [z^q B(z^{-1})C(z^{-1}) + z^{-q}B(z)C(z)],$$

where the coefficients of $C(z^{-1})$ are given in Lemma 4. The proof is completed. □

The following lemma gives the bound of the clustered eigenvalues of $Q_N^{-1}\triangle Q_N$.

LEMMA 6. *Let T_m be a sequence of $m \times m$ symmetric Toeplitz matrices generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *and let the corresponding generating sequence satisfy* (2.4) *and* (2.5). *Then $Q_N^{-1}\triangle Q_N$ has at least $N - 2\max(p,q)$ eigenvalues with magnitude bounded by*

$$\epsilon = \max_{z=e^{-i2\pi n/N}} \left| \frac{\triangle Q(z)}{Q(z)} \right| . \tag{4.5}$$

Proof. Let us denote the southeast $(N-\max(p,q)) \times (N-\max(p,q))$ blocks of Q_N and $\triangle Q_N$ by $\mathcal{Q}_{N-\max(p,q)}$ and $\triangle\mathcal{Q}_{N-\max(p,q)}$, respectively. By the minimax theorem (or *Courant–Fisher theorem*) of eigenvalues [20], [26], there are at least $N-2\max(p,q)$ eigenvalues of $Q_N^{-1}\triangle Q_N$ bounded by the maximum and the minimum eigenvalues of $\mathcal{Q}^{-1}_{N-\max(p,q)}\triangle\mathcal{Q}_{N-\max(p,q)}$.

It is clear from Lemma 5 that $\mathcal{Q}_{N-\max(p,q)}$ and $\triangle\mathcal{Q}_{N-\max(p,q)}$ are symmetric banded Toeplitz matrices with bandwidth $\leq \max(p,q)$. We construct two $N \times N$ symmetric circulant matrices $\mathcal{R}_N$ and $\triangle\mathcal{R}_N$ with $\mathcal{Q}_{N-\max(p,q)}$ and $\triangle\mathcal{Q}_{N-\max(p,q)}$ as their leading principal submatrices, respectively. By the separation theorem (or intertwining theorem) of eigenvalues [20], [26], the eigenvalues of $\mathcal{Q}^{-1}_{N-\max(p,q)}\triangle\mathcal{Q}_{N-\max(p,q)}$ are bounded by the maximum and the minimum eigenvalues of $\mathcal{R}_N^{-1}\triangle\mathcal{R}_N$. It is well known that the eigenvalues of $\mathcal{R}_N^{-1}\triangle\mathcal{R}_N$ are

$$\triangle Q(e^{-i2\pi n/N})/Q(e^{-i2\pi n/N}), \qquad n = 0, 1, \cdots, N-1.$$

Thus the proof is completed. □

We then focus on the bound of (4.5). By using (3.1) and (3.3), $\triangle Q(z)/Q(z)$ can be further simplified as

$$\begin{aligned} \triangle Q(z)/Q(z) &= [b_q z^q B(z^{-1})C(z^{-1}) + b_q z^{-q}B(z)C(z)]/[B(z^{-1})B(z)T(z)] \\ &= [b_q z^q C(z^{-1})]/[B(z)T(z)] + [b_q z^{-q}C(z)]/[B(z^{-1})T(z)]. \end{aligned} \tag{4.6}$$

Since $T(e^{i\theta}) = A(e^{-i\theta})/B(e^{-i\theta}) + A(e^{i\theta})/B(e^{i\theta})$, and $|T(e^{i\theta})|$ is finite from (2.4), $|B(e^{i\theta})|$ is uniformly positive, i.e.,

$$|B(e^{i\theta})| \geq \beta > 0. \tag{4.7}$$

Combining (2.5), (4.6), and (4.7), we obtain

$$\left|\frac{\triangle Q(e^{-i\theta})}{Q(e^{-i\theta})}\right| \leq \left|\frac{2b_q C(e^{-i\theta})}{\beta\delta}\right|, \tag{4.8}$$

with arbitrary θ.

We then focus our discussion on the bound of $|b_q C(e^{-i\theta})|$. First, we have

$$|b_q C(e^{-i\theta})| \leq \sum_{i=0}^{q} |b_q c_i| = \sum_{i=0}^{q} \left|\sum_{j=0}^{i} b_q d_j \tilde{t}'_{i-j}\right| = \sum_{i=0}^{q} \left|\sum_{j=0}^{i} b_{q-j} t_{N-i+j}\right|, \tag{4.9}$$

where the last equality is due to (3.7), (3.10a), and (3.10b). Since t_n satisfies the recursion (3.6), we use the equality

$$\sum_{j=0}^{q} b_{q-j} t_{N-i+j} = 0$$

with $N > \max(p, q)$ to simplify (4.9), i.e.,

$$\begin{aligned} |b_q C(e^{-i\theta})| &\leq \sum_{i=0}^{q} \left| - \sum_{j=i+1}^{q} b_{q-j} t_{N-i+j} \right| \leq \sum_{i=0}^{q} \sum_{j=i+1}^{q} |b_{q-j}||t_{N-i+j}| \\ &\leq \max_{N \leq n \leq N+q} |t_n| \sum_{i=0}^{q} \sum_{j=i+1}^{q} |b_{q-j}|. \end{aligned} \tag{4.10}$$

Furthermore, the term $\sum_{i=0}^{q} \sum_{j=i+1}^{q} |b_{q-j}|$ is bounded by

$$\sum_{i=0}^{q} \sum_{j=i+1}^{q} |b_{q-j}| < \sum_{i=0}^{q} \sum_{j=1}^{q} |b_{q-j}| < (q+1) \sum_{j=0}^{q} |b_j| < (q+1)2^q, \tag{4.11}$$

where the last inequality is due to the following lemma.

LEMMA 7. *Let T_m be a sequence of $m \times m$ symmetric Toeplitz matrices generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *and let the corresponding generating sequence satisfy* (2.4) *and* (2.5); *then*

$$\sum_{j=0}^{q} |b_j| < 2^q.$$

Proof. Since $B(z^{-1})$ is a polynomial in z of order q, $B(z^{-1})$ can be factorized as

$$B(z^{-1}) = \sum_{i=0}^{q} b_i z^{-i} = (1 - r_1 z^{-1})(1 - r_2 z^{-1}) \cdots (1 - r_q z^{-1}), \tag{4.12}$$

where r_i, $1 \leq i \leq q$, are poles of $T_+(z^{-1})$. A direct consequence of (2.4) is that all poles of $T_+(z^{-1})$ should lie inside the unit circle, i.e., $|r_i| < 1$, $1 \leq i \leq q$. It is clear from (4.12) that

$$|b_k| \leq \begin{pmatrix} q \\ k \end{pmatrix} (\max |r_i|)^k < \begin{pmatrix} q \\ k \end{pmatrix}, \qquad \text{where} \begin{pmatrix} q \\ k \end{pmatrix} \equiv \frac{q!}{(q-k)!\, k!}.$$

Therefore, we have

$$\sum_{j=0}^{q} |b_j| < \sum_{j=0}^{q} \binom{q}{j} = 2^q,$$

and the proof is completed. □

Combining (4.8), (4.10), and (4.11), we have

$$\max_n |\triangle Q(e^{-i2\pi n/N})/Q(e^{-i2\pi n/N})| < \frac{2^{q+1}(q+1)}{\beta\delta} \max_{N\leq n\leq N+q} |t_n|.$$

Since $|r_i| < 1$, $1 \leq i \leq q$, t_n is monotonically decreasing and

$$\max_{N\leq n\leq N+q} |t_n| = |t_N|,$$

for sufficiently large N. Thus

$$\max_n |\triangle Q(e^{-i2\pi n/N})/Q(e^{-i2\pi n/N})| < \frac{2^{q+1}(q+1)|t_N|}{\beta\delta} = \epsilon_K. \tag{4.13}$$

By Lemma 6, there are at least $N - 2\max(p,q)$ eigenvalues of $Q_N^{-1}\triangle Q_N$ with magnitude bounded by ϵ_K in (4.13). Since eigenvalues of $T_N^{-1}\triangle T_N$ are equivalent to those of $Q_N^{-1}\triangle Q_N$, there are at least $N - 2\max(p,q)$ eigenvalues of $T_N^{-1}\triangle T_N$ with magnitude bounded by ϵ_K as well. When ϵ_K is small enough, there are at least $N-2\max(p,q)$ eigenvalues of $K_{i,N}^{-1}T_N$, $i = 1,2,3,4$, clustered between $(1-\epsilon_K, 1+\epsilon_K)$ for sufficiently large N. We summarize the analysis in this section into the following theorem.

THEOREM 4. *Let T_m be a sequence of $m \times m$ symmetric Toeplitz matrices generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *and let the corresponding generating sequence satisfy* (2.4) *and* (2.5). *For sufficiently large N, the spectra of the preconditioned Toeplitz matrices $K_{i,N}^{-1}T_N$, $i = 1,2,3,4$, have the following two properties:*

P1. *The number of outliers is at most* $2\max(p,q)$.

P2. *There are at least $N-2\max(p,q)$ eigenvalues that lie between $(1-\epsilon_K, 1+\epsilon_K)$, where ϵ_K is given by* (4.13).

5. Discussion on Strang's preconditioners. We adopt a procedure similar to that described in §§3 and 4 to examine the spectral properties of $S_N^{-1}T_N$, where S_N is Strang's preconditioner. Only the cases where $p \leq q$ and $N = 2M$ are discussed. Since the analysis for the cases where $p > q$ or N is odd can be performed in a straightforward way, it is omitted to avoid unnecessary repetition.

Recall that Strang's preconditioner S_N is obtained by preserving the central half-diagonals of T_N and using them to form a circulant matrix. That is, when $N = 2M$, S_N is defined as a symmetric Toeplitz matrix with the first row

$$S_N: \quad [t_0, t_1, \cdots, t_{M-1}, t_M, t_{M-1}, \cdots, t_1].$$

Let us denote the difference between S_N and T_N by $\triangle S_N$, i.e., $\triangle S_N = S_N - T_N$. The number of outliers of $S_N^{-1}T_N$ is determined by the following lemma.

LEMMA 8. *Let T_N be an $N \times N$ symmetric Toeplitz matrix generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *and let the corresponding generating sequence satisfy* (2.4) *and* (2.5). *$T_N^{-1}\triangle S_N$ has asymptotically at most $2\max(p,q)$ nonzero eigenvalues* (*outliers*).

Proof. The proof is similar to that of Lemma 2. We use

$$\triangle E_N = \begin{bmatrix} 0 & \triangle F_M \\ \triangle F_M^T & 0 \end{bmatrix}$$

to approximate $\triangle S_N$, where

$$\triangle F_M = \begin{bmatrix} t_M & t_{M-1} & t_{M-2} & \cdots & t_3 & t_2 & t_1 \\ t_{M+1} & t_M & t_{M-1} & \cdots & \cdot & t_3 & t_2 \\ t_{M+2} & t_{M+1} & t_M & \cdots & \cdot & \cdot & t_3 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ t_{N-3} & \cdot & \cdot & \cdots & t_M & t_{M-1} & t_{M-2} \\ t_{N-2} & t_{N-3} & \cdot & \cdots & t_{M+1} & t_M & t_{M-1} \\ t_{N-1} & t_{N-2} & t_{N-3} & \cdots & t_{M+2} & t_{M+1} & t_M \end{bmatrix}.$$

Since elements t_n in $\triangle F_M$ satisfy the recursion described in Lemma 1, there are at most $\max(p,q)$ independent rows in $\triangle F_M$. Therefore, the rank of $\triangle E_N$ is at most $2\max(p,q)$. Let us define $\triangle P_N = \triangle E_N - \triangle S_N$. Then we find that

$$\triangle P_N = \begin{bmatrix} 0 & \triangle G_M \\ \triangle G_M^T & 0 \end{bmatrix},$$

where $\triangle G_M$ is an $M \times M$ symmetric Toeplitz matrix with the first row

$$\triangle G_M : \quad [t_M, t_{M+1}, t_{M+2}, \cdots, t_{N-3}, t_{N-2}, t_{N-1}].$$

It is easy to verify that, for sufficiently large N, the l_1 and l_∞ norms of $\triangle P_N$ are both less than

$$\tau_S = 2 \sum_{n=M}^{N-1} |t_n|.$$

Consequently, we have

$$||\triangle P_N||_2 \le (||\triangle P_N||_1 ||\triangle P_N||_\infty)^{1/2} \le \tau_S.$$

Since τ_S goes to zero as M goes to infinity due to (2.4), and since the eigenvalues of T_N^{-1} are bounded due to (2.5), the spectra of $T_N^{-1}\triangle S_N$ and $T_N^{-1}\triangle E_N$ are asymptotically equivalent. It follows that both $T_N^{-1}\triangle E_N$ and $T_N^{-1}\triangle S_N$ have at most $2\max(p,q)$ nonzero eigenvalues, asymptotically. □

The matrix $\triangle S_N$ can be expressed as $\triangle S_N = \triangle S_{1,N} - \triangle S_{2,N}$, where

$$\triangle S_{1,N} = \begin{bmatrix} 0 & F_{1,M} \\ F_{1,M}^T & 0 \end{bmatrix} \quad \text{and} \quad \triangle S_{2,N} = \begin{bmatrix} 0 & F_{2,M} \\ F_{2,M}^T & 0 \end{bmatrix},$$

and where $F_{1,M}$ and $F_{2,M}$ are $M \times M$ upper triangular Toeplitz matrices with the following first rows:

$$F_{1,M} : \quad [t_M, t_{M-1}, t_{M-2}, \cdots, t_2, t_1],$$
$$F_{2,M} : \quad [t_M, t_{M+1}, t_{M+2}, \cdots, t_{N-2}, t_{N-1}].$$

We use t_n, which satisfies (3.6), to construct two new sequences:

$$\tilde{s}_{1,n} = \begin{cases} 0, & 0 \leq n \leq M-1, \\ t_{N-n}, & M \leq n \leq M+q-1, \\ -(\sum_{k=1}^{q} b_{q-k}\tilde{t}_{n-k})/b_q, & M+q \leq n, \end{cases} \tag{5.1}$$
$$\tilde{s}_{2,n} = \begin{cases} 0, & 0 \leq n \leq M-1, \\ t_n, & M \leq n, \end{cases}$$

and associate $\tilde{s}_{1,n}$ and $\tilde{s}_{2,n}$ with two sequences of symmetric Toeplitz matrices $\tilde{S}_{1,m}$ and $\tilde{S}_{2,m}$, $m = 1, 2, \cdots$, whose generating functions are defined as

$$\tilde{S}_1(z) = \tilde{S}_{1,+}(z^{-1}) + \tilde{S}_{1,+}(z), \quad \text{where} \quad \tilde{S}_{1,+}(z^{-1}) = \frac{\tilde{s}_{1,0}}{2} + \sum_{n=1}^{\infty} \tilde{s}_{1,n} z^{-n},$$

and

$$\tilde{S}_2(z) = \tilde{S}_{2,+}(z^{-1}) + \tilde{S}_{2,+}(z), \quad \text{where} \quad \tilde{S}_{2,+}(z^{-1}) = \frac{\tilde{s}_{2,0}}{2} + \sum_{n=1}^{\infty} \tilde{s}_{2,n} z^{-n},$$

respectively. We can easily verify that for $N > 2\max(p,q)$,

$$\tilde{S}_{1,N} = \triangle S_{1,N} \quad \text{and} \quad \tilde{S}_{2,N} = \triangle S_{2,N}.$$

Then, by using the same approach for proving Lemma 3, we obtain the following lemma.

LEMMA 9. *If T_N is generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *then*

$$\tilde{S}_{1,+}(z^{-1}) = \frac{z^{-M}\tilde{C}(z^{-1})}{D(z^{-1})} = \frac{\tilde{c}_0 + \tilde{c}_1 z^{-1} + \cdots + \tilde{c}_q z^{-q}}{d_0 + d_1 z^{-1} + \cdots + d_q z^{-q}} \tag{5.2a}$$

and

$$\tilde{S}_{2,+}(z^{-1}) = \frac{z^{-M}\tilde{A}(z^{-1})}{B(z^{-1})} = \frac{\tilde{a}_0 + \tilde{a}_1 z^{-1} + \cdots + \tilde{a}_q z^{-q}}{b_0 + b_1 z^{-1} + \cdots + b_q z^{-q}}, \tag{5.2b}$$

where the coefficients b_i and d_i are given by (3.3) *and* (3.8), *and*

$$\tilde{a}_i = \begin{cases} \sum_{j=0}^{i} b_j t_{M+i-j}, & 0 \leq i \leq q, \\ 0, & q < i, \end{cases} \qquad \tilde{c}_i = \begin{cases} \sum_{j=0}^{i} d_j \tilde{s}_{1,M+i-j}, & 0 \leq i \leq q, \\ 0, & q < i, \end{cases}$$

with $\tilde{s}_{1,n}$ given by (5.1).

Thus $\triangle S_N$ can be decomposed into

$$\triangle S_N = \triangle S_{1,N} - \triangle S_{2,N} = L_{\tilde{c}} L_d^{-1} + U_{\tilde{c}} U_d^{-1} - L_{\tilde{a}} L_b^{-1} - U_{\tilde{a}} U_b^{-1}, \tag{5.3}$$

where $L_{\tilde{c}}$ (or $U_{\tilde{c}}$) is an $N \times N$ lower (or upper) triangular Toeplitz matrix with the first N coefficients of $z^{-M}\tilde{C}(z^{-1})$ as its first column (or row), and matrices $L_{\tilde{a}}$, L_b, and L_d (or $U_{\tilde{a}}$, U_b, and U_d) are similarly defined with respect to $z^{-M}\tilde{A}(z^{-1})$, $B(z^{-1})$, and $D(z^{-1})$, respectively.

By using the decomposition formulas (3.5) and (5.3), we transform the generalized eigenvalue problem

$$\triangle S_N \mathbf{x} = \lambda T_N \mathbf{x} \tag{5.4}$$

into another generalized eigenvalue problem

$$\Delta Q_{S,N}\mathbf{y} = \lambda Q_N \mathbf{y}, \tag{5.5}$$

where

$$\begin{aligned} Q_N = L_b T_N U_b &= L_a U_b + L_b U_a, \\ \Delta Q_{S,N} = L_b \Delta S_N U_b &= (L_b L_{\tilde c} L_d^{-1} U_b + L_b U_{\tilde c} U_d^{-1} U_b) - (L_{\tilde a} U_b + L_b U_{\tilde a}). \end{aligned}$$

The systems (5.4) and (5.5) have the same eigenvalues and their eigenvectors are related via $\mathbf{x} = U_b \mathbf{y}$. The matrix $\Delta Q_{S,N}$ is a nearly banded Toeplitz matrix characterized by the following lemma.

LEMMA 10. *Let T_m be a sequence of $m \times m$ symmetric Toeplitz matrices generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *and the corresponding generating sequence satisfies* (2.4) *and* (2.5). *The southeast $(N - \max(p,q)) \times (N - \max(p,q))$ block of $\Delta Q_{S,N}$ is a symmetric banded Toeplitz matrix with the generating function*

$$\Delta Q_S(z) = B(z^{-1})\tilde S(z)B(z) = \Delta Q_{S,1}(z) - \Delta Q_{S,2}(z), \tag{5.6}$$

where

$$\begin{aligned} \Delta Q_{S,1}(z) &= b_q z^{-(M-q)} B(z^{-1})\tilde C(z^{-1}) + b_q z^{M-q} B(z)\tilde C(z), \\ \Delta Q_{S,2}(z) &= z^{-M}\tilde A(z^{-1})B(z) + z^M B(z^{-1})\tilde A(z). \end{aligned}$$

Since the generating sequence t_n of T_N satisfies conditions (2.4) and (2.5), we can use arguments given in the previous section and obtain

$$\left| \frac{\Delta Q_{S,1}(e^{-i\theta})}{Q(e^{-i\theta})} \right| \le \frac{2^{q+1}(q+1)|t_M|}{\beta\delta} = \epsilon'$$

and

$$\left| \frac{\Delta Q_{S,2}(e^{-i\theta})}{Q(e^{-i\theta})} \right| \le \frac{2^{q+1}(q+1)|t_M|}{\beta\delta} = \epsilon'$$

for arbitrary θ. By using arguments similar to those in Lemma 6, it can be derived that $T_N^{-1}\Delta S_N$ has at least $N - 2\max(p,q)$ eigenvalues bounded by

$$\epsilon_S = 2\epsilon' = \frac{2^{q+2}(q+1)|t_{N/2}|}{\beta\delta} \tag{5.7}$$

for sufficiently large N. The analysis in this section is concluded by the following theorem.

THEOREM 5. *Let T_m be a sequence of $m \times m$ symmetric Toeplitz matrices generated by $T(z)$ with $T_+(z^{-1})$ given by* (3.3), *and the corresponding generating sequence satisfies* (2.4) *and* (2.5). *For sufficiently large N, the spectrum of the preconditioned Toeplitz matrix $S_N^{-1}T_N$ has the following two properties:*

P1. *The number of outliers is at most* $2\max(p,q)$.

P2. *There are at least $N-2\max(p,q)$ eigenvalues that lie between $(1-\epsilon_S, 1+\epsilon_S)$, where ϵ_S is given by* (5.7).

Let us compare the preconditioners $K_{i,N}$ and S_N. From Theorems 4 and 5, the spectra of $K_{i,N}^{-1}T_N$ and $S_N^{-1}T_N$ have the same number of outliers, and the other eigenvalues are clustered around 1 within radii ϵ_K and ϵ_S given by (4.13) and (5.7), respectively. It is clear that the parameters q, β, and δ are independent of the problem size N, and that the terms $|t_N|$ and $|t_{N/2}|$ determine the convergence rate of the PCG method. For sufficiently large N, we have $O(\epsilon_K) = O(\epsilon_S^2)$. This implies that, after the first several iterations which eliminate the effects of the outliers, the residual reduced by one iteration of the PCG method with preconditioners $K_{i,N}$ is about the same as that reduced by two iterations of the PCG method with preconditioner S_N. This has been confirmed by numerical experiments reported in [15].

6. The special case with geometric generating sequences. It has been observed from numerical experiments [15], [21], that the eigenvalues of $K_{1,N}^{-1}T_N$ and $S_N^{-1}T_N$ with T_N generated by the geometric sequence $t_n = t^n$, $|t| < 1$, are very regular. The observations are summarized as follows.

R1. The eigenvalues of $K_{1,N}^{-1}T_N$ are $(1+t)^{-1}$, $(1-t)^{-1}$, and $(1-t^N)^{-1}$ with multiplicities 1, 1, and $N-2$, respectively.

R2. When N is even ($N = 2M$), the eigenvalues of $S_N^{-1}T_N$ are $(1+t)^{-1}$, $(1-t)^{-1}$, 1, $(1+t^M)^{-1}$, and $(1-t^M)^{-1}$ with multiplicities 1, 1, 2, $M-2$, and $M-2$, respectively.

In this section, we provide an analytical approach to explain these two regularities.

First, we examine the preconditioner $K_{1,N}$. For the generating sequence $t_n = t^n$, its generating function is

$$T(z) = T_+(z^{-1}) + T_+(z), \quad \text{where } T_+(z^{-1}) = \frac{A(z^{-1})}{B(z^{-1})} = \frac{0.5 + 0.5tz^{-1}}{1 - tz^{-1}},$$

so that the order (p, q) of $T_+(z^{-1})$ is $(1, 1)$. From Lemma 3, we obtain

$$\tilde{T}_+(z^{-1}) = \frac{C(z^{-1})}{D(z^{-1})} = \frac{t^N(0.5 + 0.5t^{-1}z^{-1})}{1 - t^{-1}z^{-1}},$$

which is related to $\triangle T_N = K_{1,N} - T_N$. By using (4.3) and (4.4), we have

$$Q(z) = A(z^{-1})B(z) + B(z^{-1})A(z) = 1 - t^2 \tag{6.1}$$

and

$$\triangle Q(z) = -t[zB(z^{-1})C(z^{-1}) + z^{-1}B(z)C(z)] = -t^N(1 - t^2). \tag{6.2}$$

Note that $q = 1$ and $b_q = -t$ are used in deriving (6.2). Due to (6.1) and (6.2), the southeast $(N-1) \times (N-1)$ blocks of Q_N and $\triangle Q_N$ are identity matrices multiplied by the constants $1 - t^2$ and $-t^N(1 - t^2)$, respectively. Consider the following linear combination of Q_N and $\triangle Q_N$:

$$V_N = \triangle Q_N + t^N Q_N.$$

It is clear that the southeast $(N-1) \times (N-1)$ block of V_N is a zero matrix. Since the first two columns are linearly independent, and any two columns of the last $N-1$ columns of V_N are linearly dependent, V_N has a null space of dimension $N-2$. This implies that $Q_N^{-1}\triangle Q_N$, or equivalently, $T_N^{-1}\triangle T_N$, has the eigenvalue $-t^N$ with multiplicity $N-2$. Therefore, $(T_N + \triangle T_N)^{-1}T_N = K_{1,N}^{-1}T_N$ has the eigenvalue $(1-t^N)^{-1}$ with multiplicity $N-2$.

To determine the remaining two eigenvalues, i.e., the outliers, we use the technique described in [4] to transform the problem $\triangle T_N \mathbf{x} = \lambda T_N \mathbf{x}$ to another equivalent problem. Consider the case with even N ($N = 2M$). Since ΔT_N and T_N are both symmetric Toeplitz matrices, they can be expressed in the following block matrix form:

$$\Delta T_N = \begin{bmatrix} \Delta T_{1,M} & \Delta T_{2,M}^T \\ \Delta T_{2,M} & \Delta T_{1,M} \end{bmatrix} \quad \text{and} \quad T_N = \begin{bmatrix} T_{1,M} & T_{2,M}^T \\ T_{2,M} & T_{1,M} \end{bmatrix}.$$

Let W_N be the orthonormal matrix

$$W_N = \frac{1}{\sqrt{2}} \begin{bmatrix} I_M & I_M \\ -J_M & J_M \end{bmatrix},$$

where I_M and J_M are $M \times M$ identity and symmetric elementary matrices, respectively. By using the transformation

$$W_N^{-1} \Delta T_N W_N \mathbf{y} = \lambda W_N^{-1} T_N W_N \mathbf{y},$$

we obtain two decoupled subproblems,

$$(\Delta T_{1,M} - J_M \Delta T_{2,M}) \mathbf{y}_- = \lambda_- (T_{1,M} - J_M T_{2,M}) \mathbf{y}_-, \tag{6.3}$$

$$(\Delta T_{1,M} + J_M \Delta T_{2,M}) \mathbf{y}_+ = \lambda_+ (T_{1,M} + J_M T_{2,M}) \mathbf{y}_+, \tag{6.4}$$

where λ_- and λ_+ are also eigenvalues of the original problem $\Delta T_N \mathbf{x} = \lambda T_N \mathbf{x}$. Since the first rows of matrices on both sides of (6.3) are proportional by a constant $-t$, $\lambda_- = -t$ with $\mathbf{y}_- = \mathbf{e}_1$ (the unit vector with 1 at the first element) satisfies (6.3). Similarly, we can argue that $\lambda_+ = t$ with $\mathbf{y}_+ = \mathbf{e}_1$ is an eigenvalue-eigenvector pair for (6.4). Thus $1/(1-t)$ and $1/(1+t)$ are two outliers of $(T_N + \Delta T_N)^{-1} T_N = K_{1,N}^{-1} T_N$. When N is odd, the same result can be derived with a slightly modified W_N given in [4].

By using the relationship among preconditioners $K_{i,N}$, $i = 1, 2, 3, 4$, we can determine all eigenvalues of $K_{i,N}^{-1} T_N$. They all have three distinct eigenvalues (two outliers and $N-2$ clustered eigenvalues) summarized in Table 1.

TABLE 1
Eigenvalues of $K_{i,N}^{-1} T_N$.

	$K_{1,N}^{-1} T_N$	$K_{2,N}^{-1} T_N$	$K_{3,N}^{-1} T_N$	$K_{4,N}^{-1} T_N$
λ_1	$(1+t)^{-1}$	$(1+t)^{-1}$	$(1+t)^{-1}$	$(1-t)^{-1}$
λ_2	$(1-t)^{-1}$	$(1-t)^{-1}$	$(1+t^N)^{-1}$	$(1+t^N)^{-1}$
λ_3	$(1-t^N)^{-1}$	$(1+t^N)^{-1}$	$(1-t^N)^{-1}$	$(1-t^N)^{-1}$

Next, we examine Strang's preconditioner S_N with even N. When $N = 2M$, the two central rows of $S_N - T_N$ are zeros. This implies that $S_N^{-1} T_N$ has the eigenvalue 1 with multiplicity 2. By using (5.2a) and (5.2b), we have

$$\tilde{S}_{1,+}(z^{-1}) = \frac{z^{-M} \tilde{C}(z^{-1})}{D(z^{-1})} = \frac{z^{-M} t^M}{1 - t^{-1} z^{-1}},$$

$$\tilde{S}_{2,+}(z^{-1}) = \frac{z^{-M} \tilde{A}(z^{-1})}{B(z^{-1})} = \frac{z^{-M} t^M}{1 - t z^{-1}},$$

respectively. By substituting $\tilde{A}(z^{-1})$, $B(z^{-1})$, $\tilde{C}(z^{-1})$, and $D(z^{-1})$ into (5.6) and using (6.1), we obtain

$$\triangle Q_S(z) = -t^M(z^{-M} + z^M)(1 - t^2).$$

Then, the nonzero elements of $\triangle Q_{S,N-1}$, which is the southeast $(N-1)\times(N-1)$ block of $\triangle Q_{S,N}$, only occur along the $\pm M$th diagonals and take the same value $-t^M(1-t^2)$. Consider the linear combination of $\triangle Q_{S,N}$ and Q_N,

$$V_{1,N} = \triangle Q_{S,N} + t^M Q_N.$$

By adding the $k+1$th column to the $M+k+1$th column of $V_{1,N}$, for $k = 1, 2, \cdots (M-1)$, we find that the southeast $(N-1) \times (M-1)$ block of the resulting matrix is the zero matrix. Consequently, $V_{1,N}$ has a null space of dimension $M-2$ and $Q_N^{-1}\triangle Q_{S,N}$ has the eigenvalue $-t^M$ with multiplicity $M-2$. Similarly, we can show that

$$V_{2,N} = \triangle Q_{S,N} - t^M Q_N$$

has a null space of dimension $M-2$ by subtracting the $k+1$ column from the $M+k+1$th column of $V_{2,N}$, for $k = 1, 2, \cdots (M-1)$. Therefore, $Q_N^{-1}\triangle Q_{S,N}$ has the eigenvalue t^M with the same multiplicity $M-2$. As a consequence, $S_N^{-1}T_N$ has the eigenvalues $(1+t^M)^{-1}$ and $(1-t^M)^{-1}$ with multiplicity $M-2$.

To determine the remaining two eigenvalues of $S_N^{-1}T_N$, we use the same transformation discussed earlier and consider the eigenvalues of the following two subproblems:

$$(6.5) \qquad (T_{1,M} - J_M T_{2,M})\mathbf{y}_- = \lambda_-(S_{1,M} - J_M S_{2,M})\mathbf{y}_-,$$

$$(6.6) \qquad (T_{1,M} + J_M T_{2,M})\mathbf{y}_+ = \lambda_+(S_{1,M} + J_N S_{2,M})\mathbf{y}_+,$$

where $S_{1,M}$ and $S_{2,M}$ are the northwest and southwest $M \times M$ blocks of S_N, respectively. Since the first rows of matrices on both sides of (6.5) are proportional by a constant $1-t$, $\lambda_- = 1/(1-t)$ with $\mathbf{y}_- = \mathbf{e}_1$ satisfies (6.5). Similarly, $\lambda_+ = 1/(1+t)$ with $\mathbf{y}_+ = \mathbf{e}_1$ satisfies (6.6).

7. Conclusion. In this paper, we have proved the spectral properties of the preconditioned rational Toeplitz matrices $P_N^{-1}T_N$ with the preconditioner S_N proposed by Strang [19] and the preconditioners $K_{i,N}$ proposed by the authors [15]. The eigenvalues of $P_N^{-1}T_N$ are classified into two classes, i.e., the outliers and the clustered eigenvalues. The number of outliers depends on the order of the rational generating function. The clustered eigenvalues are confined in the interval $(1-\epsilon, 1+\epsilon)$ with the radii $\epsilon_K = O(|t_N|)$ and $\epsilon_S = O(|t_{N/2}|)$ for $K_{i,N}^{-1}T_N$ and $S_N^{-1}T_N$, respectively. When the symmetric Toeplitz matrix T_N is generated by the geometric sequence t^n with $|t| < 1$, the precise eigenvalue distributions of $K_{i,N}^{-1}T_N$ and $S_{2M}^{-1}T_{2M}$ have been determined analytically. Since the eigenvalues of $K_{i,N}^{-1}T_N$ are more closely clustered than those of $S_N^{-1}T_N$, preconditioners $K_{i,N}$ are more efficient for solving rational Toeplitz systems.

REFERENCES

[1] G. S. Ammar and W. B. Gragg, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[2] R. R. Bitmead and B. D. Anderson, *Asymptotically fast solution of Toeplitz and related systems of equations*, Linear Algebra Appl., 34 (1980), pp. 103–116.

[3] R. P. Brent, F. G. Gustavson, and D. Y. Yun, *Fast solution of Toeplitz systems of equations and computations of Padé approximations*, J. Algorithms, 1 (1980), pp. 259–295.

[4] A. Cantoni and P. Butler, *Eigenvalues and eigenvectors of symmetric centrosymmetric matrices*, Linear Algebra Appl., 13 (1976), pp. 275–288.

[5] R. H. Chan, *Circulant preconditioners for Hermitian Toeplitz system*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 542–550.

[6] ———, *The spectrum of a family of circulant preconditioned Toeplitz systems*, SIAM J. Numer. Anal., 26 (1989), pp. 503–506.

[7] R. H. Chan and G. Strang, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 104–119.

[8] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

[9] P. Davis, *Circulant Matrices*, John Wiley, New York, 1979.

[10] P. Delsarte and Y. V. Genin, *The split Levinson algorithm*, IEEE Trans. Acoust. Speech Signal Process., ASSP-34 (1986), pp. 470–478.

[11] B. W. Dickinson, *Efficient solution of linear equations with banded Toeplitz matrices*, IEEE Trans. Acoust. Speech Signal Process., ASSP-27 (1979), pp. 421–422.

[12] ———, *Solution of linear equations with rational Toeplitz matrices*, Math. Comp., 34 (1980), pp. 227–233.

[13] F. D. Hoog, *A new algorithm for solving Toeplitz systems of equations*, Linear Algebra Appl., 88/89 (1987), pp. 123–138.

[14] T. Huckle, *Circulant and skew-circulant matrices for solving Toeplitz matrices problems*, in Copper Mountain Conference on Iterative Methods, Copper Mountain, CO, 1990; SIAM J. Matrix Anal. Appl., 13 (1992), pp. 767–777.

[15] T. K. Ku and C. J. Kuo, *Design and analysis of Toeplitz preconditioners*, IEEE Trans. Signal Process., 40 (1992), pp. 129–141.

[16] N. Levinson, *The Wiener* RMS *error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[17] H. S. Malvar, *Fast computation of the discrete Cosine transform and the discrete Hartley transform*, IEEE Trans. Acoust. Speech Signal Process., ASSP-35 (1987), pp. 1484–1485.

[18] H. V. Sorensen, D. L. Jones, M. T. Heideman, and C. S. Burrus, *Real-value fast Fourier transform algorithms*, IEEE Trans. Acoust. Speech Signal Process., ASSP-35 (1987), pp. 849–863.

[19] G. Strang, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.

[20] ———, *Linear Algebra and Its Applications*, Harcourt Brace Jonanovich, Inc., Third Edition, Orlando, FL, 1988.

[21] G. Strang and A. Edelman, *The Toeplitz-circulant eigenvalue problem* $Ax = \lambda Cx$, in Oakland Conference on PDE's, L. Bragg and J. Dettman, eds., Longmans, London, 1987.

[22] L. N. Trefethen, *Rational approximation on the unit disk*, Numer. Math., 37 (1981), pp. 297–320.

[23] ———, *Approximation theory and numerical linear algebra*, in Algorithms for Approximation II, M. Cox and J. C. Mason, eds., Chapman, London, New York, 1988.

[24] W. F. Trench, *Solution of systems with Toeplitz matrices generated by rational functions*, Linear Algebra Appl., 74 (1986), pp. 191–211.

[25] ———, *Toeplitz systems associated with the product of a formal Laurent series and a Laurent polynomial*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 181–193.

[26] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, New York, 1965.

[27] P. Yip and K. R. Rao, *A fast computational algorithm for the discrete Sine transform*, IEEE Trans. Comm., COM-28 (1980), pp. 304–307.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 166–179, January 1993

015

BIDIRECTIONAL CHASING ALGORITHMS FOR THE EIGENVALUE PROBLEM*

DAVID S. WATKINS†

Abstract. The QR eigenvalue algorithm and its variants are usually implemented implicitly as chasing algorithms. The matrix whose eigenvalues are sought is first reduced to Hessenberg form by a similarity transformation, then the chasing iterations are begun. Each iteration is a sequence of similarity transformations that create a bulge in the Hessenberg form at one corner of the matrix, then chase the bulge along the diagonal to the opposite corner and finally off the edge of the matrix. Byers's Hamiltonian QR algorithm is an example of a chasing algorithm that has a special feature. In order to preserve the Hamiltonian structure of the matrices, the algorithm forms and chases two bulges simultaneously. Viewed in the appropriate coordinate system, this process can be seen as one in which two bulges are created at opposite corners of the matrix and chased toward each other. They collide in the middle, they interact, and then they continue to the corners of the matrix, where they vanish. In this paper it is shown that the single-shift Hamiltonian QR algorithm is just one of a family of bidirectional chasing algorithms that can be applied to arbitrary matrices. Thus the basic mechanism underlying the Hamiltonian QR algorithm does not rely on the Hamiltonian structure in any way. The structure is exploited to make the algorithm more efficient and to improve its numerical stability.

Key words. eigenvalue, QR algorithm, GR algorithm, chasing the bulge, Hamiltonian matrix

AMS(MOS) subject classifications. 65F15, 15A18

1. Introduction.

Let $A \in \mathbb{C}^{n \times n}$ be a matrix whose eigenvalues we would like to know. One of the most popular algorithms for finding the eigenvalues is the QR algorithm [6], [9], [12]. Similar algorithms named LR [12], SR [1], [3], and HR [1], [2] are also sometimes used. Recently, Watkins and Elsner [10] developed a general theory of GR algorithms that encompasses this entire family. One step of the generic GR algorithm consists of a similarity transformation

$$A_1 = G^{-1}AG,$$

where the nonsingular transforming matrix G satisfies an equation of the form

$$p(A) = GR.$$

Here p is a polynomial of low degree, and R is an upper triangular matrix. We place no conditions on G except that it be nonsingular. (However, when it comes to implementing practical algorithms, we would like our G to be well conditioned.) The simplest choice of p is $p(A) = A - \sigma I$, where σ is a *shift* chosen to approximate an eigenvalue of A. Another common choice, which yields a *double GR* step, is $p(A) = (A - \sigma I)(A - \tau I)$, where σ and τ are approximations of eigenvalues. Repeated steps of the generic GR algorithm produce a sequence (A_i) of similar matrices. If the p and G for each step are chosen in a reasonable manner, the sequence (A_i) usually converges to a block triangular form from which the eigenvalues of A can be read. For details, see [10].

* Received by the editors July 11, 1990; accepted for publication (in revised form) March 11, 1991. This article was presented at the Second SIAM Conference on Linear Algebra in Systems, Signals, and Control, San Francisco, California, November 1990. This research was supported by the National Science Foundation under grant DMS-8800437.

† Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington 99164-3113 (watkins@wsumath.bitnet).

In certain contexts, we wish to perform *GR* steps for which p is not a polynomial. For example, we can take p to be a linear fractional transformation $p(A) = (A - \sigma I)(A - \tau I)^{-1}$. If σ and τ are good approximations to distinct eigenvalues of A, then $p(A)$ has one eigenvalue that is much larger than all others and one that is much smaller, suggesting rapid convergence of certain subspace iterations that drive the *GR* algorithm (see [10]).

A specific example of an algorithm that uses a linear fractional p is Byers's single-shift Hamiltonian *QR* algorithm [5], in which A is a real Hamiltonian matrix, and $\tau = -\sigma \in \mathbb{R}$. The advantage of using this form of p is that if A is Hamiltonian, then $p(A) = (A - \sigma I)(A + \sigma I)^{-1}$ also has a special structure; it is symplectic. This permits the construction of a *QR* algorithm that preserves the Hamiltonian structure. The payoff is that arithmetic, data flow, and computer storage space are cut in half.

Byers's derivation of the Hamiltonian *QR* algorithm made heavy use of the Hamiltonian and symplectic structures. The object of this paper is to demonstrate that the basic mechanism that drives Byers's algorithm does not in any way rely on these special properties. We demonstrate here that the single-shift Hamiltonian *QR* algorithm is a special case of a class of bidirectional chasing algorithms that can act on arbitrary matrices in $\mathbb{C}^{n \times n}$.

The plan of the paper is as follows. Section 2 begins with the observation that a generic *GR* step with a linear fractional p is equivalent to a combination of a conventional *GR* step with a conventional *RG* step, in either order. We then consider how such a step can be implemented implicitly by chasing a bulge either down the diagonal and then back up, or up the diagonal and then back down. In §3, we consider bidirectional chasing steps, which chase bulges up and down the diagonal simultaneously. We prove a theorem showing that, if executed properly, the bidirectional chasing step effects one step of the generic *GR* algorithm with linear fractional p. In §4, we study Hamiltonian matrices and derive Byers's Hamiltonian *QR* step as a special case of the generic bidirectional chasing step. Of course, Byers's algorithm is the paradigm for all of the developments of this paper. In §5, we discuss briefly the difficulties we encountered when we attempted to extend the theory to cover the double-shift Hamiltonian *QR* algorithm.

2. The generic *GR* algorithm for linear fractional *p*.

In [10] we discussed the generic *GR* algorithm. We could equally well have studied the *RG* algorithm. Each step of the generic *RG* algorithm is a similarity transformation

$$A_1 = GAG^{-1},$$

where the transforming matrix G is obtained from a decomposition $q(A) = RG$, in which R is upper triangular. Again, the only condition on G is that it be nonsingular. Notice that if $q(A)$ is nonsingular, the *RG* iteration driven by $q(A)$ is equivalent to a *GR* step driven by $q(A)^{-1}$: the equation $q(A)^{-1} = G^{-1}R^{-1}$ is a *GR* decomposition, and $A_1 = (G^{-1})^{-1}AG^{-1}$.

If we wish to perform a *GR* step driven by $p(A) = (A - \sigma I)(A - \tau I)^{-1}$, we should do a *GR* step driven by $(A - \tau I)^{-1}$, followed by a *GR* step driven by $A - \sigma I$. But the former is just an *RG* step driven by $A - \tau I$, so we can carry out the step without inverting $A - \tau I$. Thus we let

$$A_{1/2} = G_{1/2}AG_{1/2}^{-1}, \tag{1}$$

where $G_{1/2}$ is from an *RG* decomposition $A - \tau I = R_{1/2}G_{1/2}$. Then we let

$$A_1 = G_1^{-1}A_{1/2}G_1, \tag{2}$$

where G_1 is from a GR decomposition $A_{1/2} - \sigma I = G_1 R_1$. To verify that these two steps do indeed constitute one GR step driven by $p(A)$, notice that

$$A_1 = G^{-1} A G,$$

where $G = G_{1/2}^{-1} G_1$, and G satisfies the equation

$$p(A) = (G_{1/2}^{-1} G_1)(R_1 R_{1/2}^{-1}) = GR.$$

We easily verify that performing the steps in the opposite order produces the same effect.[1]

Each of the half-steps (1) and (2) can be implemented by chasing. In [11] we showed how to effect a GR step by chasing. The RG step can be effected by an analogous procedure, which we outline briefly here. Recall that a matrix A is *upper Hessenberg* if $a_{ij} = 0$ for $i > j + 1$. An upper Hessenberg matrix is *irreducible* if $a_{ij} \neq 0$ for $i = j + 1$. The key theorems are Theorem 2.4 of [11] and its RG analogue.

THEOREM 2.1. *Let $A \in \mathbb{C}^{n \times n}$ be an irreducible upper Hessenberg matrix, and let p be any analytic function defined on the spectrum of A. Let G be a nonsingular matrix whose first column is proportional to $x = p(A)e_1$, such that $B = G^{-1}AG$ is upper Hessenberg. Then there exists an upper triangular matrix R such that $p(A) = GR$.*

This is essentially Theorem 2.4 of [11], except that in [11] we took p to be a polynomial. The proof carries over to analytic p without change.

THEOREM 2.2. *Let $A \in \mathbb{C}^{n \times n}$ be an irreducible upper Hessenberg matrix, and let p be any analytic function defined on the spectrum of A. Let G be a nonsingular matrix whose last row is proportional to $y^T = e_n^T p(A)$, such that $B = GAG^{-1}$ is upper Hessenberg. Then there exists an upper triangular matrix R such that $p(A) = RG$.*

The proof is essentially the same as that of Theorem 2.1.

We now outline briefly how an RG step can be implemented by chasing. Assume that A is in irreducible upper Hessenberg form. Let $\check{G}$ be a nonsingular matrix that differs from the identity matrix only in the 2×2 lower right-hand corner submatrix, whose last row is proportional to the last row of $A - \tau I$. Perform the similarity transformation $\check{G}A\check{G}^{-1}$, which creates a bulge in the upper Hessenberg form at position $(n, n-2)$. Now let $\check{G}_n$ be a matrix that differs from the identity matrix only in rows and columns $n-2$ and $n-1$, such that multiplication of $\check{G}A\check{G}^{-1}$ by $\check{G}_n^{-1}$ on the right annihilates the bulge. In other words, $\check{G}A\check{G}^{-1}\check{G}_n^{-1}$ has upper Hessenberg form. Complete the similarity transformation by multiplying by $\check{G}_n$ on the left. The resulting matrix $\check{G}_n\check{G}A\check{G}^{-1}\check{G}_n^{-1}$ has a bulge at position $(n-1, n-3)$. The bulge has been moved up and to the left. Subsequent transformations of the same type chase the bulge up the diagonal and off the edge of the matrix. The result is a matrix $A_{1/2} = G_{1/2}AG_{1/2}^{-1}$, where the last row of $G_{1/2}$ is proportional to the last row of $A - \tau I$. By Theorem 2.2 there exists an upper triangular matrix $R_{1/2}$ such that $A - \tau I = R_{1/2}G_{1/2}$. Thus the chasing step is an RG step. The GR half-step (2) can be implemented analogously, except that the bulge is formed at the top and chased to the bottom. Hence the entire step consists of chasing a bulge from bottom to top, and then chasing another bulge from top to bottom. If we decide to do the GR half-step before the RG half-step, the chasing directions are reversed.

[1] That is, we get $\tilde{A}_1 = \tilde{G}^{-1}A\tilde{G}$, where $p(A) = \tilde{G}\tilde{R}$. It is not necessarily true that $\tilde{G} = G$; this depends on which of numerous possible GR and RG decompositions are used in the half-steps.

3. Bidirectional chasing. We now consider the possibility of chasing in both directions at once. Partition A into a 2×2 block matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{11} \in \mathbb{C}^{j \times j}$. The blocks A_{11} and A_{22} are irreducible upper Hessenberg, and A_{21} consists of zeros, except for the entry $a_{j+1,j}$ in the upper right-hand corner. In particular, A_{21} has rank one. Clearly $A_{21} = a_{j+1,j}e_1e_j^T$, where $e_1 \in \mathbb{C}^{n-j}$ and $e_j \in \mathbb{C}^j$ are the usual standard unit vectors. Suppose we initiate a chasing step at the upper left-hand corner of A with a transformation whose first column is proportional to that of $A - \sigma I$, and chase the bulge to the bottom of A_{11}. At the same time we initiate a chasing step at the lower right-hand corner with a transformation whose last row is proportional to that of $A - \tau I$, and we chase the bulge to the top of A_{22}. These transformations effect a complete GR step on A_{11} and a complete RG step on A_{22}: A_{11} and A_{22} are transformed to $G_1^{-1}A_{11}G_1$ and $G_2A_{22}G_2^{-1}$, respectively, where G_1 satisfies $A_{11} - \sigma I = G_1R_1$ for some upper triangular R_1, and G_2 satisfies $A_{22} - \tau I = R_2G_2$ for some upper triangular R_2. The effect on the whole matrix A is to transform it to

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} G_1^{-1} & \\ & G_2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} G_1 & \\ & G_2^{-1} \end{bmatrix} = \begin{bmatrix} G_1^{-1}A_{11}G_1 & G_1^{-1}A_{12}G_2^{-1} \\ G_2A_{21}G_1 & G_2A_{22}G_2^{-1} \end{bmatrix}. \tag{3}$$

B_{11} and B_{22} are upper Hessenberg, and so are the transforming matrices G_1 and G_2. Since A_{21} has rank one, so does $B_{21} = G_2A_{21}G_1$. Indeed $B_{21} = a_{j+1,j}xy^T$, where $x = G_2e_1$ and $y^T = e_j^TG_1$. Since G_1 and G_2 are upper Hessenberg, only the first two entries of x and the last two entries of y are nonzero. Thus the only nonzero entries of B_{21} lie in the upper right-hand 2×2 submatrix. This small rank-one matrix is the product of the collision of the two bulges. How can we sort out the bulges and send them on their separate ways? There are two possibilities, the naive one and the one that works.

Let us consider the naive solution first and see why it fails. The Hessenberg form is disturbed by the bulge in the $(2,1)$ block; we have to get rid of it. Let $\tilde{G}_3 \in \mathbb{C}^{j \times j}$ be a matrix that differs from the identity matrix only in its lower right-hand 2×2 submatrix, whose last row is proportional to y^T. Then the similarity transformation $\operatorname{diag}\{\tilde{G}_3, I\}B\operatorname{diag}\{\tilde{G}_3^{-1}, I\}$ sets the entries $b_{j+1,j-1}$ and $b_{j+2,j-1}$ to zero and creates a bulge at position $(j, j-2)$. A sequence of similarity transformations can then chase this bulge up through the $(1,1)$ block and off the edge. Let G_3 denote the product of these similarity transformations, including $\tilde{G}_3$. The last row of G_3 is proportional to y^T.

We can chase a bulge down the $(2,2)$ block at the same time. Let $\tilde{G}_4 \in \mathbb{C}^{(n-j)\times(n-j)}$ be a matrix that differs from the identity matrix only in its upper left-hand 2×2 submatrix, whose first column is proportional to x. Then the similarity transformation $\operatorname{diag}\{I, \tilde{G}_4^{-1}\}B\operatorname{diag}\{I, \tilde{G}_4\}$ sets the entries $b_{j+2,j-1}$ and $b_{j+2,j}$ to zero and creates a bulge in position $(j+3, j+1)$. This bulge can be chased down through the $(2,2)$ block by a sequence of similarity transformations. Let G_4 denote the product of these transformations, including $\tilde{G}_4$. The first column of G_4 is proportional to x.

Applying the G_3 and G_4 transformations simultaneously, we obtain an upper Hessenberg matrix

$$C = \begin{bmatrix} G_3 & \\ & G_4^{-1} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} G_3^{-1} & \\ & G_4 \end{bmatrix}.$$

The problem with this similarity transformation is that it simply undoes the work done by the transformation from A to B. To see this it is helpful to have the following results, which are actually special cases of Theorems 2.1 and 2.2.

THEOREM 3.1. *Let $A \in \mathbb{C}^{n\times n}$ be an irreducible upper Hessenberg matrix, and let G be a nonsingular matrix whose first column is proportional to e_1, such that $B = G^{-1}AG$ is upper Hessenberg. Then G is upper triangular.*

THEOREM 3.2. *Let $A \in \mathbb{C}^{n\times n}$ be an irreducible upper Hessenberg matrix, and let G be a nonsingular matrix whose last row is proportional to e_n^T, such that $B = GAG^{-1}$ is upper Hessenberg. Then G is upper triangular.*

Proof. Theorem 3.1 (respectively, 3.2) follows from Theorem 2.1 (respectively, 2.2) by taking $p \equiv 1$. □

The complete similarity transformation from A to C is given by

$$C = \begin{bmatrix} G_3G_1^{-1} & \\ & G_4^{-1}G_2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} G_1G_3^{-1} & \\ & G_2^{-1}G_4 \end{bmatrix}. \tag{4}$$

G_1 and G_3 have proportional last rows; each is a multiple of y^T. Thus $e_j^TG_3G_1^{-1} = \alpha y^TG_1^{-1} = \beta e_j^T$ for some constants α and β. This means that the last row of $G_3G_1^{-1}$ is proportional to e_j^T. Applying Theorem 3.2, we find that $G_3G_1^{-1}$ is upper triangular. A similar relationship holds between G_2 and G_4: the first column of each is proportional to x, so the first column of $G_2^{-1}G_4$ is proportional to e_1. By Theorem 3.1, $G_2^{-1}G_4$ is upper triangular. Thus the transforming matrix in (4) is upper triangular; the similarity transformation makes no progress toward triangular form.[2] In the QR case, all of the "G" matrices are unitary, so the transforming matrix in (4) is both upper triangular and unitary. Thus it is diagonal with main diagonal entries of unit modulus, and C and A are practically identical. In the LR case without pivoting, the transforming matrix in (4) is I, so $C = A$.

Aside from the analysis, it is already evident from the form of (4) that the procedure that we have just outlined will fail: the transforming matrix in (4) is block diagonal, so there is no interaction between the top and bottom portions of the matrix. To see what might be done to remedy this, we return to the matrix B in (3). Recall that B fails to be upper Hessenberg only in that B_{21} has four nonzero entries in its upper right-hand corner instead of one. These form a 2×2 block of rank one. Before we return the matrix to upper Hessenberg form, we must perform some kind of similarity transformation that mixes the top and bottom blocks. It is easy to check that a transformation that recombines only rows (and columns) j and $j+1$ does not alter the pattern of zeros in B. In fact, this is the only type of similarity transformation that combines rows from the upper and lower blocks without disturbing the zero pattern. Let S be a matrix that performs such a similarity transformation. Then S

[2] The entire transformation can be viewed as a step of the GR algorithm driven by $p(A) = I$, for $C = G^{-1}AG$, where $G = \operatorname{diag}\{G_1G_3^{-1}, G_2^{-1}G_4\}$ satisfies the GR decompositon $p(A) = I = GR$ with the upper triangular R given by $R = G^{-1}$. Thus the underlying subspace iterations (driven by $p(A) = I$) are stationary.

has the form

$$\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & s_{11} & s_{12} & & \\ & & & s_{21} & s_{22} & & \\ & & & & & 1 & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{bmatrix},$$

where s_{11} is the (j,j) entry. We call a matrix of this form a *mixing matrix.* The submatrix

$$\tilde{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$$

is called the *essential submatrix* of the mixing matrix S. The matrix $T = S^{-1}$ is also a mixing matrix. Its essential submatrix

$$\tilde{T} = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}$$

is the inverse of $\tilde{S}$.

We wish to examine the effect on B of a similarity transformation by a mixing matrix. In particular, we are interested in the effect on the 2×2 submatrix in the upper right-hand corner of B_{21}. For typographic simplicity we let

$$\begin{bmatrix} \alpha & \gamma \\ \beta & \delta \end{bmatrix}$$

denote this submatrix. The transformation $S^{-1}BS$ changes it to

$$(5) \quad \begin{bmatrix} t_{21}b_{j,j-1} + t_{22}\alpha & (t_{21}b_{j,j} + t_{22}\gamma)s_{11} + (t_{21}b_{j,j+1} + t_{22}b_{j+1,j+1})s_{21} \\ \beta & \delta s_{11} + b_{j+2,j+1}s_{21} \end{bmatrix}.$$

If we wish to return the resulting matrix to upper Hessenberg form by transformations that do not mix the upper and lower blocks, we must choose S so that (5) has rank one, for its rank will not be altered by the subsequent transformations. Setting the determinant of (5) to zero and using the equations $\alpha\delta - \beta\gamma = 0$, $t_{21} = -s_{21}/d$, and $t_{22} = s_{11}/d$, where d is the determinant of S, we obtain

$$s_{21}\left\{s_{11}[\beta(b_{j,j} - b_{j+1,j+1}) + \alpha b_{j+2,j+1} - b_{j,j-1}\delta] - s_{21}[b_{j,j-1}b_{j+2,j+1} - \beta b_{j,j+1}]\right\} = 0.$$

The solution $s_{21} = 0$ is the "naive" one; it just rescales the rows and columns of B_{21}. The subsequent reduction to upper Hessenberg form essentially undoes the first part of the transformation. Thus the only solution that might be of any use is

$$(6) \quad s_{11}[\beta(b_{j,j} - b_{j+1,j+1}) + \alpha b_{j+2,j+1} - b_{j,j-1}\delta] = s_{21}[b_{j,j-1}b_{j+2,j+1} - \beta b_{j,j+1}].$$

There are many mixing matrices S that satisfy this property, but they all have essentially the same effect. (We ignore the issue of numerical stability for now.) All that

TABLE 1
Bidirectional chasing step.

```
do simultaneously
    choose shift $\sigma$
    construct $\hat{G}$ for which $\hat{G}e_1 \propto (A - \sigma)e_1$
    $A \leftarrow \hat{G}^{-1}A\hat{G}$        (create bulge at $a_{31}$)
    for $i = 1, 2, \ldots, j - 2$
        $A \leftarrow \hat{G}_i^{-1}A\hat{G}_i$        (chase bulge from $a_{i+2,i}$ to $a_{i+3,i+1}$)
and
    choose shift $\tau$
    construct $\check{G}$ for which $e_n^T\check{G} \propto e_n^T(A - \tau)$
    $A \leftarrow \check{G}A\check{G}^{-1}$        (create bulge at $a_{n,n-2}$)
    for $i = n, n - 1, \ldots, j + 3$
        $A \leftarrow \check{G}_iA\check{G}_i^{-1}$        (chase bulge from $a_{i,i-2}$ to $a_{i-1,i-3}$)
end do
construct mixing matrix $S$ satisfying
    $s_{11}[a_{j+2,j-1}(a_{jj} - a_{j+1,j+1}) + a_{j+1,j-1}a_{j+2,j+1} - a_{j,j-1}a_{j+2,j}]$
        $= s_{21}[a_{j,j-1}a_{j+2,j+1} - a_{j+2,j-1}a_{j,j+1}]$
$A \leftarrow S^{-1}AS$
do simultaneously
    for $i = j + 1, j, \ldots, 3$
        $A \leftarrow \check{G}_iA\check{G}_i^{-1}$        (chase bulge from $a_{i,i-2}$ to $a_{i-1,i-3}$)
and
    for $i = j, j + 1, \ldots, n - 2$
        $A \leftarrow \hat{G}_i^{-1}A\hat{G}_i$        (chase bulge from $a_{i+2,i}$ to $a_{i+3,i+1}$)
end do
```

matters is the proportion $s_{21} : s_{11}$, for this determines the proportions in which the jth and $(j+1)$st rows (and columns) are combined to form the new $(j+1)$st row (and jth column). Suppose we perform a transformation of this form and then reduce the resulting matrix to upper Hessenberg form by a similarity transformation of the form

$$\begin{bmatrix} G_3^{-1} & \\ & G_4 \end{bmatrix},$$

as before. Then the combined transforming matrix has the form

$$G = \begin{bmatrix} G_1 & \\ & G_2^{-1} \end{bmatrix} S \begin{bmatrix} G_3^{-1} & \\ & G_4 \end{bmatrix}.$$

$G_1, \ldots, G_4$ are upper Hessenberg matrices, and S is a mixing matrix whose entries satisfy (6). As we shall see, this transformation is the right one; that is, it effects a step of the generic GR algorithm driven by $(A - \sigma I)(A - \tau I)^{-1}$. The key result to this end is Theorem 3.3, which is stated and proved below.

The bidirectional chasing step is summarized in Table 1. Each of the transformations used in the chasing step consists of a nonsingular 2×2 essential submatrix embedded in what would otherwise be an $n \times n$ identity matrix. In order to perform the tasks in this algorithm, we need essential submatrices whose first column or last row is proportional to any prescribed vector in $\mathbb{C}^2$.[3] The mixing matrix does not

[3] Notice that specifying the direction of the first column of a 2×2 matrix $\tilde{S}$ is the same as specifying the direction of the last row of $\tilde{S}^{-1}$, since the last row of $\tilde{S}^{-1}$ is orthogonal to the first column of $\tilde{S}$.

differ in form from the bulge-chasing matrices. In general, we allow the use of any nonsingular matrices that can perform the required tasks. If we wish to restrict our transformations to some subclass of the nonsingular matrices (e.g., unitary matrices), we need only ensure that the subclass contains matrices of the form just stated above. In §4, when we consider bidirectional chasing algorithms that preserve Hamiltonian structure, we impose one additional requirement: our mixing matrices must have determinant 1. Finally, in practice, stability and convergence considerations dictate that the transforming matrices be well conditioned.

We now state and prove the key result.

THEOREM 3.3. *Let $A \in \mathbb{C}^{n\times n}$ be in irreducible upper Hessenberg form, and suppose σ and τ are distinct complex numbers that are not eigenvalues of A. Let j be an integer satisfying $1 \leq j < n$, and consider the partition*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $A_{11} \in \mathbb{C}^{j\times j}$. Let $G_1, R_1 \in \mathbb{C}^{j\times j}$ and $G_2, R_2 \in \mathbb{C}^{(n-j)\times(n-j)}$ be nonsingular matrices, R_1 and R_2 upper triangular, satisfying

$$A_{11} - \sigma I = G_1 R_1, \qquad A_{22} - \tau I = R_2 G_2.$$

(This implies that G_1 and G_2 are irreducible upper Hessenberg.) Let S be any mixing matrix satisfying (6)*, where B is as in* (3)*. Then there exist G, $R \in \mathbb{C}^{n\times n}$ such that*

$$(A - \sigma I)(A - \tau I)^{-1} = GR,$$

R is upper triangular, and G has the form

$$G = \begin{bmatrix} G_1 & \\ & G_2^{-1} \end{bmatrix} S \begin{bmatrix} G_3^{-1} & \\ & G_4 \end{bmatrix}, \tag{7}$$

where G_3 and G_4 are upper Hessenberg.

Proof. We will construct a matrix G of the stated form and another nonsingular matrix Z such that $U_1 = G^{-1}(A-\sigma I)Z$ and $U_2 = (A-\tau I)Z$ are both upper triangular. Then, letting $R = U_1 U_2^{-1}$, we have $(A-\sigma I)(A-\tau I)^{-1} = GR$, with R upper triangular.

We need to reduce $A-\tau I$ to upper triangular form by multiplication on the right. First of all,

$$(A - \tau I)\mathrm{diag}\{I, G_2^{-1}\} = \begin{bmatrix} A_{11} - \tau I & X \\ A_{21} & R_2 \end{bmatrix}.$$

Here and throughout the proof, the symbol X in the $(1,2)$ block denotes a matrix whose entries are of no immediate interest. The value of X can change from one occurrence to the next. R_2 is upper triangular, and A_{21} consists entirely of zeros, except for the entry $a_{j+1,j}$ in the upper right-hand corner. We wish to eliminate $a_{j+1,j}$. Let r denote the $(1,1)$ entry of R_2, let

$$\tilde{V} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$$

be any nonsingular matrix for which

$$\begin{bmatrix} a_{j+1,j} & r \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix} = \begin{bmatrix} 0 & * \end{bmatrix}, \tag{8}$$

and let $V \in \mathbb{C}^{n\times n}$ be the mixing matrix whose essential submatrix is $\tilde{V}$. Then

$$(A-\tau I)\mathrm{diag}\{I, G_2^{-1}\}V = \begin{bmatrix} \tilde{H}_{11} & X \\ 0 & \tilde{R}_2 \end{bmatrix},$$

where $\tilde{R}_2$ is upper triangular, and $\tilde{H}_{11}$ is upper Hessenberg.

Finally, let F_1 be a nonsingular matrix such that $\tilde{H}_{11}F_1$ is upper triangular. Letting $\tilde{R}_1 = \tilde{H}_{11}F_1$, we note that $F_1^{-1} = \tilde{R}_1^{-1}\tilde{H}_{11}$, so F_1^{-1} is upper Hessenberg. Let $Z = \mathrm{diag}\{I, G_2^{-1}\}V\mathrm{diag}\{F_1, I\}$. Then

$$(A-\tau I)Z = \begin{bmatrix} \tilde{R}_1 & X \\ 0 & \tilde{R}_2 \end{bmatrix},$$

which is upper triangular.

Now we have to find G, of the stated form, for which $G^{-1}(A-\sigma I)Z$ is upper triangular. First notice that

$$\mathrm{diag}\{G_1^{-1}, G_2\}(A-\sigma I)\mathrm{diag}\{I, G_2^{-1}\} = \begin{bmatrix} R_1 & X \\ G_2A_{21} & G_2A_{22}G_2^{-1} - \sigma I \end{bmatrix}.$$

R_1 is upper triangular, and $G_2A_{22}G_2^{-1} - \sigma I = G_2R_2 + (\tau - \sigma)I$ is upper Hessenberg. Letting g denote the $(2,1)$ entry of G_2, and recalling that the $(1,1)$ entry of R_2 is called r, we see that the $(2,1)$ entry of $G_2R_2 + (\tau-\sigma)I$ is gr. The block G_2A_{21} has zeros everywhere except positions $(1,j)$ and $(2,j)$. The $(2,j)$ entry is $ga_{j+1,j}$. Since

$$\begin{bmatrix} ga_{j+1,j} & gr \end{bmatrix}\begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix} = \begin{bmatrix} 0 & * \end{bmatrix},$$

by (8), we can set the $(2,j)$ entry of G_2A_{21} to zero by multiplying by V on the right. That is,

$$\mathrm{diag}\{G_1^{-1}, G_2\}(A-\sigma I)\mathrm{diag}\{I, G_2^{-1}\}V = \begin{bmatrix} \check{R}_1 & X \\ \check{A}_{21} & \check{H}_{22} \end{bmatrix},$$

where $\check{R}_1$ is upper triangular, $\check{H}_{22}$ is upper Hessenberg, and $\check{A}_{21}$ consists of zeros, except for an entry $\check{a}_{j+1,j}$ in the upper right-hand corner.

The next task is to eliminate $\check{a}_{j+1,j}$. Let $\check{r}$ denote the (j,j) entry of $\check{R}_1$. Let $T = S^{-1}$, where S is the mixing matrix named in the statement of the theorem. As we shall see, the essential submatrix of T satisfies

$$\begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}\begin{bmatrix} \check{r} \\ \check{a}_{j+1,j} \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix}. \tag{9}$$

Indeed, this is so if and only if $t_{21}\check{r} + t_{22}\check{a}_{j+1,j} = 0$, which is equivalent to

$$s_{11}\check{a}_{j+1,j} = s_{21}\check{r}. \tag{10}$$

This equation determines the proportion $s_{21} : s_{11}$. We will see below that this is exactly the proportion specified by (6), so the essential submatrix of T satisfies (9). Accepting for now that this is true, we have

$$S^{-1}\begin{bmatrix} \check{R}_1 & X \\ \check{A}_{21} & \check{H}_{22} \end{bmatrix} = \begin{bmatrix} \hat{R}_1 & X \\ 0 & \hat{H}_{22} \end{bmatrix},$$

where $\hat{R}_1$ is upper triangular, and $\hat{H}_{22}$ is upper Hessenberg.

Let G_3 be any nonsingular matrix for which $G_3\hat{R}_1F_1$ is upper triangular. Calling this product R_{11}, we have $G_3 = R_{11}F_1^{-1}\hat{R}_1^{-1}$. Since F_1^{-1} is upper Hessenberg, so is G_3. Let G_4 be any nonsingular matrix for which $G_4^{-1}\hat{H}_{22}$ is upper triangular. Calling this matrix R_{22}, we have $G_4 = \hat{H}_{22}R_{22}^{-1}$, so G_4 is upper Hessenberg. By a straightforward computation we have

$$\begin{bmatrix} G_3 & \\ & G_4^{-1} \end{bmatrix} \begin{bmatrix} \hat{R}_1 & X \\ 0 & \hat{H}_{22} \end{bmatrix} \begin{bmatrix} F_1 & \\ & I \end{bmatrix} = \begin{bmatrix} R_{11} & X \\ 0 & R_{22} \end{bmatrix},$$

which is upper triangular. Call this matrix U_1. Letting

$$G = \begin{bmatrix} G_1 & \\ & G_2^{-1} \end{bmatrix} S \begin{bmatrix} G_3^{-1} & \\ & G_4 \end{bmatrix},$$

we easily compute that $G^{-1}(A-\sigma I)Z = U_1$.

The proof will be complete as soon as we show that (10) is (6). We begin by noting that the construction that we have just outlined can be carried through with S replaced by any mixing matrix $\hat{S}$ whose inverse T satisfies (9), which is equivalent to (10). There are many matrices that satisfy this property, and for any such $\hat{S}$, the resulting G satisfies $(A-\sigma I)(A-\tau I)^{-1} = GR$ for some upper triangular R. Let $C = G^{-1}AG$. Then, as $G = p(A)R^{-1}$, $C = RAR^{-1}$, so C is upper Hessenberg. Let

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} = \hat{S}^{-1} \begin{bmatrix} G_1^{-1} & \\ & G_2 \end{bmatrix} A \begin{bmatrix} G_1 & \\ & G_2^{-1} \end{bmatrix} \hat{S} = \hat{S}^{-1}B\hat{S}.$$

Then K has the same zero pattern as B. Also, the rank of K_{21} is 1, for the similarity transformation $C = \operatorname{diag}\{G_3, G_4^{-1}\}K\operatorname{diag}\{G_3^{-1}, G_4\}$ implies $K_{21} = G_4C_{21}G_3$, and C_{21} has rank one. Thus $\hat{S}$ must satisfy either (6) or $\hat{s}_{21} = 0$. We show that the latter is impossible. If $\hat{s}_{21} = 0$, then G is block upper triangular, which implies that $p(A) = (A-\sigma I)(A-\tau I)^{-1}$ is also block upper triangular. Since $\sigma \neq \tau$, the linear fractional transformation p has an inverse q, and $A = q(p(A))$. Thus A is also block upper triangular. But this contradicts the fact that A has irreducible upper Hessenberg form. Therefore $\hat{s}_{21} \neq 0$, and $\hat{S}$ must satisfy (6). Since $\hat{S}$ also satisfies (10), we see that (6) and (10) are the same. □

The main theorem follows directly from Theorem 3.3.

THEOREM 3.4. *Under the same assumptions and terminology as in Theorem* 3.3, *let* $G_3' \in \mathbb{C}^{j\times j}$ *and* $G_4' \in \mathbb{C}^{(n-j)\times(n-j)}$ *be any nonsingular upper Hessenberg matrices for which* $C' = G'^{-1}AG'$ *is upper Hessenberg, where*

$$G' = \begin{bmatrix} G_1 & \\ & G_2^{-1} \end{bmatrix} S \begin{bmatrix} G_3'^{-1} & \\ & G_4' \end{bmatrix}.$$

Then there is an upper triangular matrix R' *such that* $(A-\sigma I)(A-\tau I)^{-1} = G'R'$. *In other words, the similarity transformation from* A *to* C' *performs one step of the generic GR algorithm driven by* $p(A) = (A-\sigma I)(A-\tau I)^{-1}$.

Proof. Let G_3, G_4, and G be as in Theorem 3.3, and let $C = G^{-1}AG$. In the proof of Theorem 3.3 we noted that C is upper Hessenberg, since $C = RAR^{-1}$. This equation implies that it is even irreducible upper Hessenberg. G' and G are related by $G' = GV$, where

$$V = G^{-1}G' = \begin{bmatrix} G_3G_3'^{-1} & \\ & G_4^{-1}G_4' \end{bmatrix}.$$

We will show that V is upper triangular. Let $V_1 = G_3'G_3^{-1}$ and $V_2 = G_4^{-1}G_4'$, so that $V = \mathrm{diag}\{V_1^{-1}, V_2\}$. Then, as $C = VC'V^{-1}$, $C_{21} = V_2C_{21}'V_1$. Since $C_{21} = \alpha e_1e_j^T$ and $C_{21}' = \alpha'e_1e_j^T$, where α and α' are scalars, we have $\alpha e_1e_j^T = \alpha'(V_2e_1)(e_j^TV_1)$, so that

$$e_j^TV_1 = \gamma e_j^T \qquad \text{and} \qquad V_2e_1 = \beta e_1$$

for certain scalars β and γ. Since also

$$C_{11}' = V_1C_{11}V_1^{-1} \qquad \text{and} \qquad C_{22}' = V_2^{-1}C_{22}V_2,$$

C_{11} and C_{22} are irreducible upper Hessenberg, and C_{11}' and C_{22}' are upper Hessenberg, we can conclude from Theorems 3.2 and 3.1 that both V_1 and V_2 are upper triangular. Thus V is upper triangular. We know from Theorem 3.3 that G satisfies $(A-\sigma I)(A-\tau I)^{-1} = GR$ for some upper triangular R. Let R' be the upper triangular matrix $V^{-1}R$. Then $(A-\sigma I)(A-\tau I)^{-1} = G'R'$. □

4. Byers's algorithm. Our final task is to show that Byers's single-shift Hamiltonian QR algorithm is a special case of the generic bidirectional chasing algorithm described in the previous section. We begin by recalling some definitions.

We now restrict our attention to real matrices, and we take n to be even, say $n = 2m$. Define $J \in \mathbb{R}^{n\times n}$ by

$$J = \begin{bmatrix} 0 & I_m \\ -I_m & 0 \end{bmatrix}.$$

A matrix $A \in \mathbb{R}^{n\times n}$ is *Hamiltonian* if $(JA)^T = JA$. Equivalently, A is Hamiltonian if and only if it has the form

$$A = \begin{bmatrix} F & N \\ K & -F^T \end{bmatrix}, \tag{11}$$

where $K^T = K$ and $N^T = N$. The eigenvalue problem for Hamiltonian matrices is important because it arises in the solution of the quadratic regulator problem of linear system theory [7].

In solving the Hamiltonian eigenvalue problem by a chasing algorithm or any algorithm that uses similarity transformations, it is desirable to preserve the Hamiltonian structure. A class of transforming matrices that achieves this end is the group of symplectic matrices. A matrix $S \in \mathbb{R}^{n\times n}$ is *symplectic* if $S^TJS = J$. We easily show that if H is Hamiltonian and S is symplectic, then $S^{-1}HS$ is Hamiltonian.

Byers's algorithm and the generalizations that we consider make use of two different types of symplectic transformations. First, if $G \in \mathbb{R}^{m\times m}$ is any nonsingular matrix, then

$$\begin{bmatrix} G & 0 \\ 0 & G^{-T} \end{bmatrix}$$

is symplectic. In particular, if $P \in \mathbb{R}^{m\times m}$ is orthogonal, then $\mathrm{diag}\{P, P\}$ is both orthogonal and symplectic. This type of symplectic transformation, when applied to a Hamiltonian matrix (11), transforms the blocks separately. We need a class of mixing matrices to mix up the blocks. First note that in the 2×2 case a matrix

$$\tilde{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$$

is symplectic if and only if $\det(\tilde{S}) = 1$. More generally, the matrix

$$S = \begin{bmatrix} I_{m-1} & & & \\ & s_{11} & & s_{12} \\ & & I_{m-1} & \\ & s_{21} & & s_{22} \end{bmatrix} \tag{12}$$

is symplectic if and only if $\det(\tilde{S}) = 1$. In particular, if $\tilde{S}$ is taken to be a rotator

$$\begin{bmatrix} c & -s \\ s & c \end{bmatrix} \qquad (c^2 + s^2 = 1),$$

S is both orthogonal and symplectic. There are obvious generalizations of this construction, but this is all we need.

We need a Hamiltonian version of upper Hessenberg form. Paige and Van Loan [8] found a method of reducing a Hamiltonian matrix by finitely many orthogonal symplectic similarity transformations to a form

$$M = \begin{bmatrix} H & N \\ D & -H^T \end{bmatrix},$$

where H is upper Hessenberg (thus $-H^T$ is lower Hessenberg) and D is diagonal. Of course $N = N^T$ as well. If the underlying control problem has a single input, then D can be taken to have the even sparser form $D = \mathrm{diag}\{0, \ldots, 0, d_m\}$. We call this form *Hamiltonian Hessenberg form*. It is on this form that Byers's algorithm operates. Notice that the Hamiltonian Hessenberg form can be converted to a true Hessenberg form by the similarity transformation that reverses the last m rows and columns of the matrix.

We describe a generalization of Byers's single-shift algorithm. In order to make the connection with bidirectional chasing algorithms, we describe the algorithm in the coordinate system in which the last m rows and columns are reversed. Let $P \in I\!R^{m\times m}$ denote the permutation matrix that reverses rows or columns. Instead of operating on the Hamiltonian Hessenberg matrix M, we operate on the true Hessenberg matrix

$$\mathrm{diag}\{I, P\} M \mathrm{diag}\{I, P\} = \begin{bmatrix} H & NP \\ PD & -PH^TP \end{bmatrix}.$$

Our algorithm is just the bidirectional chasing algorithm applied to this matrix, but we put some constraints on how the step is carried out. First of all, the shifts σ and τ are chosen to be real numbers such that $\tau = -\sigma$. Thus we begin by forming two bulges and chasing them to the middle of the matrix. This portion of the algorithm performs a similarity transformation by a matrix $\mathrm{diag}\{G_1, G_2^{-1}\}$ such that

$$H - \sigma I = G_1 R_1 \quad \text{and} \quad -PH^TP + \sigma I = R_2 G_2,$$

where R_1 and R_2 are upper triangular. It is possible to do this in such a way that G_1 and G_2 have a simple relationship: the equation $H - \sigma I = G_1 R_1$ implies that

$$-PH^TP + \sigma I = (-PR_1^TP)(PG_1^TP).$$

Since $-PR_1^TP$ is upper triangular, we can (and will) take $G_2 = PG_1^TP$. Thus the transforming matrix is $\mathrm{diag}\{G_1, PG_1^{-T}P\}$. The equivalent transforming matrix in the original coordinate system is $\operatorname{diag}\{G_1, G_1^{-T}\}$, which is symplectic. Thus the Hamiltonian structure is preserved in this portion of the step. In carrying out the step, only one of the two bulges needs to be chased in practice, since the two bulge-chasing procedures are essentially identical.

Our matrix has now been transformed to the form (3). The next step is to apply a mixing matrix S that satisfies (6). In addition we require that the essential submatrix $\tilde{S}$ satisfies $\det(\tilde{S}) = 1$. This is always possible; we can take $\tilde{S}$ to be a rotator, for example. The equivalent transforming matrix in the original coordinate system has the form (12), so it is symplectic. Thus this step also preserves the Hamiltonian structure.

At this point our matrix has the form

$$\begin{bmatrix} \hat{H} & \hat{N}P \\ P\hat{D} & -P\hat{H}^TP \end{bmatrix}.$$

The block $P\hat{D}$ has nonzeros only in the 2×2 patch in the upper right-hand corner, and it has rank one. Furthermore, $\hat{D}$ is symmetric, so $\hat{D} = \alpha vv^T$ for some $\alpha \in I\!R$ and $v \in I\!R^m$ that has nonzeros only in its last two components.

The step is continued by annihilating three of the nonzero entries of $P\hat{D}$, leaving only a single nonzero in the upper right-hand corner. This creates bulges in $\hat{H}$ and $-P\hat{H}^TP$, which we then chase out to the edge of the matrix to complete the bidirectional chasing step.

The net effect of this second chasing phase is to apply a similarity transformation of the form $\operatorname{diag}\{G_3^{-1}, G_4\}$ such that $G_3\hat{H}G_3^{-1}$ and $-G_4^{-1}P\hat{H}^TPG_4$ are upper Hessenberg. Since $P\hat{D} = \alpha(Pv)v^T$, G_3 and G_4 must also satisfy $e_m^TG_3 = \beta v^T$ and $G_4e_1 = \gamma Pv$ for some β and γ. Once again, it is possible to perform these operations in such a way that G_3 and G_4 are related. First, note that the equation $e_m^TG_3 = \beta v^T$ implies that $(PG_3^TP)e_1 = \beta Pv$, which suggests that it might be possible to take $G_4 = PG_3^TP$. This is confirmed by noting that with this choice of G_4, we have $-G_4^{-1}P\hat{H}^TPG_4 = -P(G_3\hat{H}G_3^{-1})^TP$, which is upper Hessenberg. We make this choice of G_4. Thus the transforming matrix for this part of the step is $\operatorname{diag}\{G_3^{-1}, PG_3^TP\}$. The equivalent transforming matrix in the original coordinate system is $\operatorname{diag}\{G_3^{-1}, G_3^T\}$, which is symplectic. Thus the Hamiltonian form is still preserved. In carrying out the step, it is again the case that only one of the two bulges has to be chased, since the two bulge-chasing procedures are again essentially identical.

Byers's Hamiltonian QR algorithm is the special case in which all of the transforming matrices are taken to be orthogonal. This is the best choice from the standpoint of stability, but there are other less expensive options that might also work well. For example, the G_i can be built up by Gaussian elimination with partial pivoting, giving a Hamiltonian LR algorithm. Also possible are hybrid algorithms that mix Gaussian elimination transformations with orthogonal transformations. The convergence of all of these algorithms is governed by the theorems in [10], which state that if the condition numbers of the accumulated transformation matrices stay bounded and the shifts converge, then the algorithm converges. If the generalized Rayleigh-quotient shift strategy is used (in this case $\sigma = -h_{11}$), the local convergence rate is quadratic.

5. Problems with the double-shift algorithm. Byers also developed a double-shift algorithm for use with complex conjugate shifts [4]. If H is a real Hamiltonian matrix and σ is a complex shift, then $(H - \sigma I)(H - \bar{\sigma}I)(H + \sigma I)^{-1}(H + \bar{\sigma}I)^{-1}$ is a real symplectic matrix that can be used to drive a double Hamiltonian QR step that uses only real arithmetic. If we wish to develop a bidirectional chasing algorithm that generalizes this double step, we must consider GR steps driven by matrices of the form $(A - \sigma_1 I)(A - \sigma_2 I)(A - \tau_1 I)^{-1}(A - \tau_2 I)^{-1}$. Such a GR step is equivalent to a (double) GR step driven by $(A - \sigma_1 I)(A - \sigma_2 I)$ followed by a (double) RG step driven by $(A - \tau_1 I)(A - \tau_2 I)$, or vice versa. This can be implemented by chasing a double bulge down the diagonal and back up again or up the diagonal and back down again. If one wishes to chases bulges in both directions at once, one must consider what happens when the bulges collide. Now the resulting rank-one matrix is not 2 × 2 but 3 × 3, and the mixing matrices can have essential submatrices as big as 4 × 4. We were unable to determine how the mixing matrix should be chosen so as to obtain the desired outcome. In an effort to gain some insight into the nature of the mixing matrix, we attempted to prove an analogue of Theorem 3.3, but we failed there as well. We have no doubt that such algorithms and theorems exist and await discovery by a sufficiently insightful investigator.

REFERENCES

[1] W. Bunse and A. Bunse-Gerstner, *Numerische lineare Algebra,* Teubner, Stuttgart, 1985.

[2] A. Bunse-Gerstner, *An analysis of the HR algorithm for computing the eigenvalues of a matrix,* Linear Algebra Appl., 35 (1981), pp. 155–178.

[3] A. Bunse-Gerstner and V. Mehrmann, *A symplectic QR-like algorithm for the solution of the real algebraic Riccati equation,* IEEE Trans. Automat. Control, AC-31 (1986), pp. 1104–1113.

[4] R. Byers, *Hamiltonian and Symplectic Algorithms for the Algebraic Riccati Equation,* Ph.D. thesis, Cornell University, Ithaca, NY, 1983.

[5] R. Byers, *A Hamiltonian QR algorithm,* SIAM J. Sci. Statist. Comput., 7 (1986), pp. 212–229.

[6] G. Golub and C. Van Loan, *Matrix Computations,* Second Edition, The Johns Hopkins University Press, Baltimore, MD, 1989.

[7] T. Kailath, *Linear Systems,* Prentice–Hall, Englewood Cliffs, NJ, 1980.

[8] C. Paige and C. Van Loan, *A Schur decomposition for Hamiltonian matrices,* Linear Algebra Appl., 41 (1981), pp. 11–32.

[9] D. S. Watkins, *Fundamentals of Matrix Computations,* John Wiley, New York, 1991.

[10] D. S. Watkins and L. Elsner, *Convergence of algorithms of decomposition type for the eigenvalue problem,* Linear Algebra Appl., 143 (1991), pp. 19–47.

[11] ———, *Chasing algorithms for the eigenvalue problem,* SIAM J. Matrix Anal. Appl., 12 (1991), pp. 374–384.

[12] J. H. Wilkinson, *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford, U.K., 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 180–194, January 1993

016

DEFERRED SHIFTING SCHEMES FOR PARALLEL QR METHODS*

ROBERT A. VAN DE GEIJN[†]

Abstract. Parallel implementation of the QR algorithm for solving the symmetric eigenvalue problem requires more than a straightforward transcription of sequential code to parallel code. Experimental adjustments include new shifting techniques and omission of the initial reducing to upper Hessenberg form. In this paper, the theory of the convergence of algorithms of decomposition type for the algebraic eigenvalue problem developed by Watkins and Elsner is generalized. The results are extended to deferred shifting schemes, which allow pipelining of iterations in parallel implementations, and analyzing the deterioration of the convergence rate. Furthermore, it is shown that eigenvalues need not be simple to obtain quadratic convergence for nondefective nonsymmetric matrices and cubic convergence for symmetric matrices.

Key words. convergence, eigenvalue, parallel computing, QR algorithm, Schur decomposition

AMS(MOS) subject classifications. 65F15, 15A18

1. Introduction.

QR algorithms have traditionally been the algorithms of choice for finding all eigenvalues of dense, tridiagonal, and upper Hessenberg matrices. The main objective of this paper is to discuss shifting techniques for such algorithms that can be used to increase the efficiency of implementations of these algorithms on concurrent computers. In addition, we extend some earlier theoretical results for general (QR-like) algorithms of decomposition type, such as the QR and LR algorithms.

With the advent of parallel computers, the preferred status of QR algorithms for the symmetric eigenvalue problem has had to be reevaluated. For tridiagonal symmetric matrices, divide and conquer techniques [4] and bisection methods [2], [8] have proven to parallelize more conveniently, yielding algorithms that have been reported to outperform the symmetric QR algorithm, even on sequential computers. For dense symmetric matrices, parallel Jacobi methods [3], [11] are competitive with parallel versions of the above-mentioned methods, which have to be preceded by a reduction to tridiagonal form.

Efficient parallel implementation of algorithms for the nonsymmetric eigenvalue problem has been more difficult. Jacobi-like algorithms have been developed by Stewart [13] and Eberlein [5], however, convergence of these methods is not guaranteed. Efficiency of fine-grained, parallel, nonsymmetric QR algorithms is hampered by the fact that while iterations of unshifted QR algorithms can be pipelined, inclusion of a shifting scheme is necessary to speed convergence [15]. Since the data used to calculate the shifts are among the last to be computed during an iteration, the addition of a shifting strategy hinders pipelining of iterations. Efficiency of coarser-grained algorithms is hindered by the amount of communication that is necessary to broadcast transformations [16]. The effects of this can be reduced by simultaneously performing several iterations with different shifts.

* Received by the editors April 16, 1990; accepted for publication (in revised form) March 21, 1991. This research was supported in part by Air Force Office of Scientific Research grant AFOSR-82-0078 and by National Science Foundation Science and Technology Center Cooperative Agreement No. CCR-8809615.

[†] Department of Computer Sciences, The University of Texas, Austin, Texas 78712 (rvdg@cs.utexas.edu). This research is based in part on a dissertation written by the author while at the University of Maryland, College Park, Maryland, under the direction of Professor G. W. Stewart. Final revisions were made while the author was visiting the University of Tennessee, Knoxville, Tennessee.

One approach that has been successfully used for improving efficiency on vector computers and shared memory multiprocessors generalizes techniques used for the double-shifted implicit QR algorithms by implicitly performing several iterations associated with multiple shifts known as *generalized Rayleigh-quotient shifts* [1], [18]. In this paper, we propose and analyze a *deferred* shifting strategy that allows pipelining of iterations by using "old" shifts, i.e., shifts from an iteration other than the current one. Convergence is shown to be slower than that of traditional shifting strategies, yet may be acceptable under some circumstances.

It is not clear that fine-grained parallel implementations of the QR algorithm will start by reducing the matrix to upper Hessenberg form, as is customary for sequential algorithms. One consequence of the reduction to upper Hessenberg form is that if the eigenvalues of the original matrix are nondefective, all eigenvalues of the upper Hessenberg matrix can be assumed to be simple, since otherwise the problem can be deflated. As a result, analyses that show the convergence rates to be quadratic and cubic for nonsymmetric and symmetric matrices, respectively, have been able to assume that all eigenvalues are simple. To date, no efficient fine-grained reduction to upper Hessenberg form has been discovered. It is not clear that a large number of processors can be usefully utilized during subsequent iterations, even if the matrix is first reduced. As a result, it is conceivable that fine-grained parallel QR algorithms will skip this reduction. One such example can be found in [15]. This prompted the question of how convergence is affected by the presence of multiple eigenvalues.

In §2, we informally examine the effects of deferring the Rayleigh-quotient shift. Section 3 formally generalizes the results of Watkins and Elsner [17] to include the deferred shift. In addition, we show how some conditions required in their results can be relaxed. In §4, we discuss the deterioration of the convergence rate when the shift is deferred. Results from numerical experiments are reported in §5. Conclusions are given in §6. It should be noted that §3 may be bypassed without loss of understanding of the concepts.

2. Deferred Rayleigh-quotient shift. In this section, we informally examine how the convergence rate of the shifted QR algorithm is affected by deferring the shift. We only consider the special case where the deferred QR algorithm is applied to a real upper Hessenberg matrix with real eigenvalues. In the next section, a thorough analysis is given for more general deferred shifting schemes and a generalization of the QR algorithm.

Given an $n \times n$ matrix A, there exists unitary matrix Q and upper triangular matrix R so that $A = QR$. This decomposition, known as the QR decomposition of A, forms the basis for the QR algorithm, which is given by Algorithm 2.1.

ALGORITHM 2.1. QR Algorithm

```
$A_0 = A$
for $i = 0, 1, \ldots$
    $A_i - s_i I \rightarrow Q_i R_i$   (QR decomposition)
    $A_{i+1} \leftarrow R_i Q_i + s_i I (= Q_i^{\mathrm{T}} A_i Q_i)$
```

The algorithm is typically preceded by a reduction to upper Hessenberg form. Under mild conditions, A_i converges to a block upper triangular matrix. The shifting sequence $\{s_i\}$ can be chosen to speed the converge [6]. A simple choice is the Rayleigh-quotient shift, $s_i = a_{nn}^{(i)}$, the (n, n) element of the ith iteration. This shifting strategy is generalized to the *deferred Rayleigh-quotient shift* by taking $s_i = a_{nn}^{(i-h)}$, where $h \geq 0$, i.e., the shift is deferred for h iterations.

Assume that the matrix A, whose eigenvalues are to be computed, is of upper Hessenberg form. This automatically implies that all iterates of the QR algorithm are of upper Hessenberg form. Following the techniques in [14], assume that the QR decomposition of the ith iteration in Algorithm 2.1 has proceeded to the point where

$$A_i - s_i I = Q_i^* R_i^* = Q_i^* \begin{pmatrix} R_{11}^{(i)} & R_{12}^{(i)} \\ 0 & A_{22}^{*(i)} \end{pmatrix},$$

where $R_{11}^{(i)}$ is upper triangular, Q_i^* is unitary, $A_{22}^{*(i)}$ is the 2×2 matrix

$$A_{22}^{*(i)} = \begin{pmatrix} a & b \\ a_{n(n-1)}^{(i)} & a_{nn}^{(i)} - s_i \end{pmatrix},$$

a is "large," and $a_{n(n-1)}^{(i)}$ and $(a_{nn}^{(i)} - s_i)$ are "small." The QR decomposition is completed by annihilating the $(n, n-1)$ element:

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} a & b \\ a_{n(n-1)}^{(i)} & a_{nn}^{(k)} - s_i \end{pmatrix} = \begin{pmatrix} r_{(n-1)(n-1)} & r_{(n-1)n} \\ 0 & r_{nn} \end{pmatrix},$$

where

$$c = a \Big/ \sqrt{a^2 + a_{n(n-1)}^{(i)2}} \quad \text{and} \quad s = a_{n(n-1)}^{(i)} \Big/ \sqrt{a^2 + a_{n(n-1)}^{(i)2}}.$$

It can be easily verified that after the formation of A_{i+1},

$$a_{n(n-1)}^{(i+1)} = \frac{(a_{nn}^{(i)} - s_i)a - (a_{n(n-1)}^{(i)} b) a_{n(n-1)}^{(i)}}{a^2 + a_{n(n-1)}^{(i)2}}$$

and

$$(1) \qquad |a_{n(n-1)}^{(i+1)}| \le \frac{|a_{nn}^{(i)} - s_i||a_{n(n-1)}^{(i)}|}{|a|} + O(|a_{n(n-1)}^{(i)}|^2).$$

If the shift equals the nondeferred Rayleigh-quotient shift, the first term is zero, implying that the convergence rate is quadratic.

Assume the Rayleigh-quotient shift is deferred for h iterations, where $h > 0$, so that $s_i = a_{nn}^{(i-h)}$. Assuming the method converges, $a_{n(n-1)}^{(i)} \to 0$ and $a_{nn}^{(i)} \to \lambda \in \lambda(A)$. Hence $|a_{nn}^{(i)} - s_i| = |a_{nn}^{(i)} - a_{nn}^{(i-h)}|$ in (1) converges to zero and the method is superlinearly convergent. Moreover, under the same assumption, $|a_{n(n-1)}^{(i+1)}| < |a_{n(n-1)}^{(i)}|$ for large i.

We now examine how quickly $a_{n(n-1)}^{(i)}$ converges to zero by bounding $|\lambda - a_{nn}^{(i)}|$.

LEMMA 2.2. *If λ is a simple eigenvalue of A and $a_{n(n-1)}^{(i)} \to 0$, then for large enough i,*

$$|a_{nn}^{(i)} - \lambda| \le C|a_{n(n-1)}^{(i)}|,$$

where C depends on A and λ.

Proof. Partition

$$A_i = \begin{pmatrix} A_{11}^{(i)} & b_{12}^{(i)} \\ b_{21}^{(i)\mathrm{T}} & a_{nn}^{(i)} \end{pmatrix}.$$

Without loss of generality, assume that $\lambda \neq a_{nn}^{(i)}$. Furthermore, assume that $A_{11}^{(i)} - \lambda I$ is nonsingular. Then there exists a nonzero vector $x = (x_1^{\mathrm{T}}, x_2^{\mathrm{T}})^{\mathrm{T}}$, such that

$$\begin{pmatrix} A_{11}^{(i)} - \lambda I & b_{12}^{(i)} \\ b_{21}^{(i)\mathrm{T}} & a_{nn}^{(i)} - \lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where x_2 is a nonzero scalar. Hence

$$x_1 = -x_2(A_{11}^{(i)} - \lambda I)^{-1} b_{12}^{(i)},$$
$$(a_{nn}^{(i)} - \lambda)x_2 = -b_{21}^{\mathrm{T}} x_1,$$

and

$$a_{nn}^{(i)} - \lambda = b_{21}^{(i)\mathrm{T}} (A_{11}^{(i)} - \lambda I)^{-1} b_{12}^{(i)}.$$

If $||(A_{11}^{(i)} - \lambda I)^{-1}||_2 \leq C_1$ for all i sufficiently large, Lemma 2.2 follows with $C = C_1||A||_2$. Clearly, $||b_{21}^{(i)}||_2 = |a_{n(n-1)}^{(i)}|$ if $A^{(i)}$ is upper Hessenberg. The condition that the norm of $(A_{11}^{(i)} - \lambda I)^{-1}$ must be bounded for large i is automatically satisfied if the method converges and λ is a simple eigenvalue, as we shall see in the next section. □

It follows that for large i,

$$|a_{nn}^{(i)} - a_{nn}^{(i-h)}| \leq |a_{nn}^{(i)} - \lambda| + |a_{nn}^{(i-h)} - \lambda| \leq C|a_{n(n-1)}^{(i)}| + C|a_{n(n-1)}^{(i-h)}| \leq 2C|a_{n(n-1)}^{(i-h)}|.$$

Combining (1) with Lemma 2.2 and the above equation yields Theorem 2.3.

THEOREM 2.3. *Under the conditions of Lemma* 2.2, *there exists a constant* $K = K(A)$ *such that for large* i,

$$|a_{n(n-1)}^{(i+1)}| \leq K|a_{n(n-1)}^{(i-h)}||a_{n(n-1)}^{(i)}|.$$

3. Generalizations. We now treat the deferred shifting scheme more formally. To do so, we adopt the notation and approach of Watkins and Elsner [17].

The QR algorithm is generalized in [17] as follows.

ALGORITHM 3.1. GR Algorithm

```
$A_0 = A$
for $i = 0, 1, \ldots$
    $p_{i+1}(A_i) \rightarrow G_i R_i$
    $A_{i+1} \leftarrow G_i^{-1} A_i G_i$
```

where G_i and R_i are nonsingular and upper triangular, respectively, and p_{i+1} is a polynomial. Since $A_i = \hat{G}_i^{-1} A_0 \hat{G}_i$, $\lambda(A_i) = \lambda(A_0)$ for all i, where $\hat{G}_i = G_0 G_1 \cdots G_{i-1}$. For the QR algorithm, all G_i are unitary.

We will concentrate on GR algorithms that use *deferred generalized Rayleigh-quotient shifts*, which compute p_{i+1} to equal the characteristic polynomial of $A_{22}^{(i-h)}$.

Here h is the number of iterations by which a shift is deferred and A_i is partitioned like

$$A_i = \begin{pmatrix} A_{11}^{(i)} & A_{12}^{(i)} \\ A_{21}^{(i)} & A_{22}^{(i)} \end{pmatrix}, \tag{2}$$

$A_{11}^{(i)} \in \mathbf{C}^{k\times k}$. For simplicity, p_{i+1} is computed to equal the characteristic polynomial of $A_{22}^{(0)}$ when $i < h$.

DEFINITION 3.2. Let $\mathcal{S}$ and $\mathcal{T}$ be subspaces. Then the *distance* between them, $d(\mathcal{S},\mathcal{T})$, equals $\|P_{\mathcal{S}} - P_{\mathcal{T}}\|_2$, where $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$ are orthogonal projections onto $\mathcal{S}$ and $\mathcal{T}$, respectively.

The following theorem generalizes and formalizes the results in the previous section for the GR algorithm.

THEOREM 3.3. *Let $\{A_i\}$ be the sequence generated by Algorithm* 3.1 *using deferred generalized Rayleigh-quotient shifts with polynomials of degree $m = n-k$ and deferring parameter h. Let $\{\lambda_1, \dots \lambda_n\}$ be the eigenvalues of A_0, partitioned into two disjoint sets, $\Lambda_1 = \{\lambda_1, \dots, \lambda_k\}$ and $\Lambda_2 = \{\lambda_{k+1}, \dots, \lambda_n\}$, and let the eigenvalues in Λ_2 be nondefective. Define $\mathcal{T}$ to be the invariant subspace corresponding to Λ_1 and let $\mathcal{S}_i$ be the subspace spanned by the first k columns of $\hat{G}_i$. Assume $\kappa(\hat{G}_i) \le \hat{\kappa}$, some positive constant. Then there exist $\epsilon > 0$ and $C > 0$ so that if $d(\mathcal{S}_i,\mathcal{T}) \le \epsilon, 0 \le i \le h$, then $d(\mathcal{S}_i,\mathcal{T}) \to 0$, and*

$$d(\mathcal{S}_{i+h+1},\mathcal{T}) \le C d(\mathcal{S}_{i+h},\mathcal{T}) d(\mathcal{S}_i,\mathcal{T}). \tag{3}$$

If, in addition, one or both of the following conditions are true:

(i) $\|A_{12}^{(i)}\|_2 = \|A_{21}^{(i)}\|_2$ *and $A_{22}^{(i)}$ is normal for all i, or*

(ii) $\|A_{12}^{(i)}\|_2 = \|A_{21}^{(i)}\|_2$ *and all eigenvalues of A are simple,*

then if ϵ is small enough,

$$d(\mathcal{S}_{i+h+1},\mathcal{T}) \le C d(\mathcal{S}_{i+h},\mathcal{T}) d(\mathcal{S}_i,\mathcal{T})^2. \tag{4}$$

Here $\kappa(\hat{G}_i)$ denotes the condition number of $\hat{G}_i$.

Note that in the case of the QR algorithm $\kappa(\hat{G}_i) = 1$ for all i. Also, if the QR algorithm is applied to a symmetric matrix, condition (i) is met.

This theorem generalizes Watkins and Elsner's Theorem 6.3 [17] as follows:

- In the case of general or normal matrices, some of the eigenvalues of A are only required to be nondefective as opposed to all having to be distinct.
- Deferred shifting is included.

Furthermore, we choose to formulate the theorem as a local convergence theorem, since the condition that the starting matrix must be close to convergence ($d(\mathcal{S}_i,\mathcal{T}) < \epsilon$) can typically be met by first iterating using an unshifted GR algorithm, i.e., by taking $p_{i+1}(A_i) = A_i$ and setting $A_0 = A_i$ once the iteration has converged sufficiently. In practice, shifting from the start typically leads to convergence. Precise conditions can be derived from [17].

The following key lemma allows us to relax the restriction made in [17] requiring the eigenvalues to be distinct. It is essentially a generalization of Lemma 2.2.

LEMMA 3.4. *Let $A, G \in \mathbf{C}^{n\times n}$ and*

$$B = G^{-1}AG = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where $B_{11} \in \mathbf{C}^{k\times k}$, $\kappa_F(G) \le \hat{\kappa}$. *Let* $\lambda(B_{11}) = \{\mu_1, \dots, \mu_k\}$ *and* $\lambda(B_{22}) = \{\mu_{k+1}, \dots, \mu_n\}$. *Then there exists* $\epsilon = \epsilon(A, \hat{\kappa})$ *such that* $||B_{21}||_2 \le \epsilon$ *implies that* $\lambda(A) = \{\lambda_i\}$ *can be ordered so that*

$$|\lambda_i - \mu_i| < \frac{\delta}{2}, \qquad 1 \le i \le n, \tag{5}$$

where

$$\delta = \min_{\lambda\in\lambda(A)} \min_{\substack{\mu \in \lambda(\mathrm{A}) \\ \mu \ne \lambda}} |\lambda - \mu|.$$

If $\{\lambda_1, \dots, \lambda_k\} \cap \{\lambda_{k+1}, \dots, \lambda_n\} = \emptyset$ *and* $\lambda_{k+1}, \dots, \lambda_n$ *are nondefective eigenvalues, then there exists a constant* $C = C(A, \hat{\kappa})$ *such that*

$$|\lambda_i - \mu_i| \le C||B_{21}||_2, \qquad k < i \le n. \tag{6}$$

If, in addition, either of the following conditions are true:

(i) $||B_{12}||_2 = ||B_{21}||_2$ *and* B_{22} *is normal, or*

(ii) $||B_{12}||_2 = ||B_{21}||_2$ *and all eigenvalues of* A *are simple,*

then there exist $\bar{\epsilon} = \bar{\epsilon}(A, \hat{\kappa}) \le \epsilon$ *and* $\bar{C} = \bar{C}(A, \hat{\kappa})$ *such that* $||B_{21}||_2 \le \bar{\epsilon}$ *implies that*

$$|\lambda_i - \mu_i| \le \bar{C}||B_{21}||_2^2, \qquad k < i \le n \tag{7}$$

for the same ordering of the eigenvalues.

Proof. Let $Q^H BQ = D + N$ be a Schur decomposition of B, where D is diagonal and N is strictly upper triangular. By Theorem 7.2.3 in [6], if $\mu \in \lambda(B + E)$ and p equals the smallest positive integer such that $N^p = 0$, then

$$\min_{\lambda\in\lambda(B)} |\lambda - \mu| \le \max(\theta, \theta^{1/p}),$$

where

$$\theta = ||E||_2 \sum_{i=0}^{p-1} ||N||_2^i.$$

Since $||N||_2 \le ||N||_F \le ||B||_F \le \kappa_F(G)||A||_F$, and $p \le n$,

$$\theta \le ||E||_2 \hat{\kappa}^{n-1} \sum_{i=0}^{n-1} ||A||_F^i.$$

Clearly, $\epsilon = \epsilon(A, \hat{\kappa})$ can be chosen so that $||E||_2 \le \epsilon$ implies that $\max(\theta, \theta^{1/p}) < \delta/2$. A continuity argument similar to the one used for Gershgorin discs [19] shows that the eigenvalues $\{\lambda_i\}$ of A can actually be paired with $\{\mu_i\}$, the eigenvalues of the perturbed matrix $B + E$, so that $|\lambda_i - \mu_i| < \delta/2$. Taking

$$E = \begin{pmatrix} 0 & 0 \\ -B_{21} & 0 \end{pmatrix}$$

completes the proof of (5).

For the second part, assume that $\lambda \in \{\lambda_{k+1}, \dots, \lambda_n\}$ is nondefective and has multiplicity m. There exists a unitary matrix $Q_{22} \in \mathbf{C}^{(n-k)\times(n-k)}$ such that $Q_{22}^{\mathrm{H}} B_{22} Q_{22} =$

$D_{22}+N_{22}$, where D_{22} is diagonal and N_{22} is strictly upper triangular. Moreover, the diagonal elements of D_{22} can be assumed to be ordered so that the m eigenvalues of B_{22} closest to λ equal the last m elements on the diagonal of D_{22}. Then

$$\begin{pmatrix} I & 0 \\ 0 & Q_{22}^{\mathrm{H}} \end{pmatrix} B \begin{pmatrix} I & 0 \\ 0 & Q_{22} \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12}Q_{22} \\ Q_{22}^{\mathrm{H}}B_{21} & D_{22}+N_{22} \end{pmatrix} \equiv \bar{B}.$$

Repartition

$$\bar{B} = \begin{pmatrix} \bar{B}_{11} & \bar{B}_{12} \\ \bar{B}_{21} & \bar{B}_{22} \end{pmatrix},$$

where $\bar{B}_{22} \in \mathbf{C}^{m\times m}$. Clearly, $||\bar{B}_{21}||_2 \leq ||B_{21}||_2 \leq \epsilon$.

By the first part of the proof of this lemma, with

$$E = \begin{pmatrix} 0 & 0 \\ -\bar{B}_{21} & 0 \end{pmatrix},$$

we know there exists an ordering of $\{\bar{\mu}_i\}$, the combined eigenvalues of $\lambda(\bar{B}_{11})$ and $\lambda(\bar{B}_{22})$, such that $|\lambda_i-\bar{\mu}_i| < \delta/2$. However, by the first part of this lemma, $\bar{\mu} \in \lambda(\bar{B}_{22})$ implies that $|\bar{\mu}-\lambda| < \delta/2$, since the eigenvalues of $\bar{B}_{22}$ equal the eigenvalues of B_{22} closest to λ. Also, $|\bar{\mu}-\lambda| > \delta/2$ for all $\bar{\mu} \in \lambda(\bar{B}_{11})$.

If $\bar{Q}_{11}^{\mathrm{H}}\bar{B}_{11}\bar{Q}_{11} = \bar{D}_{11}+\bar{N}_{11}$, then

$$\begin{aligned} ||(\bar{B}_{11}-\lambda)^{-1}||_2 &\leq ||(\bar{D}_{11}-\lambda I)^{-1}||_2||(I+(\bar{D}_{11}-\lambda I)^{-1}\bar{N}_{11})^{-1}||_2 \\ &\leq \frac{2}{\delta}\sum_{i=0}^{n-m-1}\left(\frac{2}{\delta}||\bar{N}_{11}||_2\right)^i \leq \frac{2}{\delta}\sum_{i=0}^{n-m-1}\left(\frac{2}{\delta}\kappa_F(G)||A||_F\right)^i \qquad (8)\\ &\leq \frac{2}{\delta}\max\left(1,\left(\frac{2}{\delta}\hat{\kappa}\right)^{n-m-1}\right)\sum_{i=0}^{n-m-1}||A||_F^i. \end{aligned}$$

Since λ is a nondefective eigenvalue of $\bar{B}$, $\mathrm{rank}(\bar{B}_{11}-\lambda I) = \mathrm{rank}(\bar{B}-\lambda I) = n-m$ and there exists a matrix $Y \in \mathbf{C}^{m\times(n-m)}$ such that

$$Y(\bar{B}_{11}-\lambda I \quad \bar{B}_{12}) = (\bar{B}_{21} \quad \bar{B}_{22}-\lambda I).$$

Hence $\bar{B}_{22}-\lambda I = \bar{B}_{21}(\bar{B}_{11}-\lambda I)^{-1}\bar{B}_{12}$. As a result, $\mu \in \lambda(\bar{B}_{22})$ implies that

$$|\mu-\lambda| \leq ||\bar{B}_{22}-\lambda I||_2 \leq ||\bar{B}_{21}||_2||(\bar{B}_{11}-\lambda I)^{-1}||_2||\bar{B}_{12}||_2 \leq C||B_{21}||_2, \qquad (9)$$

where

$$C = \frac{2}{\delta}||A||_2\max\left(1,\left(\frac{2}{\delta}\hat{\kappa}\right)^{n-1}\right)||A||_2\sum_{i=0}^{n-1}||A||_F^i.$$

Since λ can be any element of $\{\lambda_{k+1},\ldots,\lambda_n\}$, (6) follows.

If condition (i) of Lemma 3.4 is met, then $N_{22}=0$ and (7) follows from (9), with $\bar{C}=2/\delta$. The proof when condition (ii) holds follows from Lemma 6.4 and the proof of Theorem 6.5 in [17]. □

Lemma 3.4 allows us to bound the difference between the eigenvalues of the submatrix from which a shifting polynomial is computed and the eigenvalues of the original matrix. In the context of the GR algorithm, the bound equals a constant multiple of $||A_{21}^{(i)}||_2$ provided the original eigenvalue is not defective and the method is already close to convergence.

The proof of Theorem 3.3 invokes several results from [17]. For the reader's convenience, we now state the relevant results from that paper.

Lemma 3.5 relates the distance of the subspaces after coordinates have been changed to the distance between the original subspaces.

LEMMA 3.5. *Let $\mathcal{S}$ and $\mathcal{T}$ be two subspaces of $\mathbb{C}^n$ of the same dimension, let $V \in \mathbf{C}^{n\times n}$ be nonsingular, and let $\tilde{\mathcal{S}} = V^{-1}\mathcal{S}$ and $\tilde{\mathcal{T}} = V^{-1}\mathcal{T}$. Then $d(\mathcal{S},\mathcal{T}) \leq \kappa_2(V)d(\tilde{\mathcal{S}},\tilde{\mathcal{T}})$.*

The next lemma indicates the effect of one step of a shifted subspace iteration when the matrix is of special form, namely block diagonal.

LEMMA 3.6. *Let $\tilde{\mathcal{T}} = \langle e_1,\ldots,e_k\rangle$ and $\tilde{\mathcal{U}} = \langle e_{k+1},\ldots,e_n\rangle$, let $\tilde{\mathcal{S}} \in \mathbf{C}^n$ be a k-dimensional space such that*

$$\tilde{\mathcal{S}} \cap \tilde{\mathcal{U}} = \{0\}, \tag{10}$$

and let $\beta = d(\tilde{\mathcal{S}},\tilde{\mathcal{T}}) < 1$. Let $T \in \mathbf{C}^{n\times n}$ be a block diagonal matrix $T = \mathrm{diag}(T_1,T_2)$, where $T_1 \in \mathbf{C}^{k\times k}$; let p be a polynomial such that $p(T_1)$ is nonsingular; and let $\tilde{\mathcal{S}}' = p(T)\tilde{\mathcal{S}}$. Then

$$d(\tilde{\mathcal{S}}',\tilde{\mathcal{T}}) \leq \frac{\beta}{\sqrt{1-\beta^2}}||p(T_2)||_2||p(T_1)^{-1}||_2.$$

Conditions under which (10) holds are given by the following lemma.

LEMMA 3.7. *Let $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{T}}$ be subspaces of $\mathbb{C}^n$ of the same dimension, and let $\tilde{\mathcal{U}}$ be the orthogonal complement of $\tilde{\mathcal{T}}$. Then $\tilde{\mathcal{S}} \cap \tilde{\mathcal{U}} = \{0\}$ if and only if $d(\tilde{\mathcal{S}},\tilde{\mathcal{T}}) < 1$.*

One final technical lemma links the convergence of the subspace $\mathcal{S}_i$ to the convergence of iterates A_i to upper triangular form.

LEMMA 3.8. *Let $A \in \mathbf{C}^{n\times n}$, and let $\mathcal{T}$ be a k-dimensional subspace of $\mathbf{C}^n$ which is invariant under A. Let $G \in \mathbf{C}^{n\times n}$ be a nonsingular matrix, and let $\mathcal{S}$ be the space spanned by the first k columns of G. Let $B = G^{-1}AG$ and consider the partitioning*

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where $B_{11} \in \mathbf{C}^{k\times k}$. Then

$$||B_{21}||_2 \leq 2\sqrt{2}\kappa_2(G)||A||_2 d(\mathcal{S},\mathcal{T}).$$

We are now ready to prove the main result.

Proof of Theorem 3.3. We first show that if ϵ is small enough, $d(\mathcal{S}_i,\mathcal{T}) \leq \epsilon$ and $d(\mathcal{S}_{i+h},\mathcal{T}) \leq \epsilon$ implies that

$$d(\mathcal{S}_{i+h+1},\mathcal{T}) \leq Cd(\mathcal{S}_{i+h},\mathcal{T})d(\mathcal{S}_i,\mathcal{T}). \tag{11}$$

Restrictions on ϵ are indicated throughout the proof.

Since $p_{i+h+1}(A_{i+h}) = G_{i+h}R_{i+h}$,

$$p_{i+h+1}(A)\hat{G}_{i+h} = \hat{G}_{i+h+1}R_{i+h}$$

and

$$\mathcal{S}_{i+h+1} = p_{i+h+1}(A)\mathcal{S}_{i+h}.$$

Let $A = X^{-1}JX$ be a Jordan decomposition, $J = \mathrm{diag}(T_1, T_2)$, where

$$T_1 = \mathrm{diag}(J_{m_1}(\lambda_{i_1}), \ldots, J_{m_{\bar{k}}}(\lambda_{i_{\bar{k}}})) \quad \text{and} \quad T_2 = \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_n).$$

Let $\tilde{\mathcal{S}}_i = X^{-1}\mathcal{S}_i$, $\tilde{\mathcal{T}} = X^{-1}\mathcal{T} = \langle e_1, \ldots, e_k \rangle$, and $\tilde{\mathcal{U}} = \langle e_{k+1}, \ldots, e_n \rangle$. Then

$$\tilde{\mathcal{S}}_{i+h+1} = p_{i+h+1}(J)\tilde{\mathcal{S}}_{i+h}.$$

Lemma 3.5 implies that $d(\tilde{\mathcal{S}}_{i+h}, \tilde{\mathcal{T}}) \leq \kappa(X) d(\mathcal{S}_{i+h}, \mathcal{T}) \leq \kappa(X)\epsilon$ and hence if ϵ is small enough, $\beta = d(\tilde{\mathcal{S}}_{i+h}, \tilde{\mathcal{T}}) \leq \sqrt{3}/2$. This implies that $\tilde{\mathcal{S}}_{i+h} \cap \tilde{\mathcal{U}} = \{0\}$ by Lemma 3.7, and thus we can apply Lemma 3.6 to obtain

$$\begin{aligned} d(\tilde{\mathcal{S}}_{i+h+1}, \tilde{\mathcal{T}}) &\leq \frac{\beta}{\sqrt{1-\beta^2}} ||p_{i+h+1}(T_2)||_2 ||p_{i+h+1}(T_1)^{-1}||_2 \\ &\leq 2\kappa(X) d(\mathcal{S}_{i+h}, \mathcal{T}) ||p_{i+h+1}(T_2)||_2 ||p_{i+h+1}(T_1)^{-1}||_2. \end{aligned} \tag{12}$$

Since T_2 is diagonal, we can bound

$$||p_{i+h+1}(T_2)||_2 \leq \max_{k<j\leq n} ||p_{i+h+1}(\lambda_j)||_2.$$

Note that $p_{i+h+1}(\lambda_j) = \prod_{l=k+1}^{n} (\lambda_j - \mu_l^{(i)})$, where $\lambda(A_{11})^{(i)} = \{\mu_j^{(i)}\}_{j=1}^{k}$ and $\lambda(A_{22}^{(i)}) = \{\mu_j^{(i)}\}_{j=k+1}^{n}$.

In order to bound $|\lambda_j - \mu_l^{(i)}|$, $k < l \leq n$, note that $A_i = \hat{G}_i^{-1} A_0 \hat{G}_i$ and hence Lemma 3.8 implies that

$$||A_{21}^{(i)}||_2 \leq 2\sqrt{2}\kappa_2(\hat{G}_i) ||A_0||_2 d(\mathcal{S}_i, \mathcal{T}) \leq 2\sqrt{2}\hat{\kappa}_2 ||A_0||_2 d(\mathcal{S}_i, \mathcal{T}).$$

Defining δ as in Lemma 3.4, if ϵ is small enough, Lemma 3.4 implies that the eigenvalues of $A_{11}^{(i)}$ and $A_{22}^{(i)}$ can be ordered so that

$$|\lambda_j - \mu_j^{(i)}| < \frac{\delta}{2}, \tag{13}$$

and there exists a constant C_1 such that

$$|\lambda_l - \mu_l^{(i)}| \leq C_1 ||A_{21}^{(i)}|| \leq C_1 2\sqrt{2}\hat{\kappa}_2 ||A||_2 d(\mathcal{S}_i, \mathcal{T}), \qquad k < l \leq n. \tag{14}$$

Since $|\lambda_j - \mu_l^{(i)}| \leq |\lambda_j| + |\mu_l^{(i)}| \leq |\lambda_j| + |\lambda_l| + |\lambda_l - \mu_l^{(i)}|$,

$$|\lambda_j - \mu_l^{(i)}| \leq 3||A_0||_2$$

if $\epsilon \leq 1/(C_1 2\sqrt{2}\hat{\kappa}_2)$, in which case

$$||p_{i+h+1}(T_2)||_2 \leq (3||A||_2)^{m-1} C_1 2\sqrt{2}\hat{\kappa}_2 ||A||_2 d(\mathcal{S}_i, \mathcal{T}). \tag{15}$$

Finally, we bound

$$\begin{aligned} ||p_{i+1}(T_1)^{-1}||_2 &\leq \max_{0<j\leq \bar{k}} ||p_{i+h+1}(J_{m_j}(\lambda_j))^{-1}||_2 \\ &\leq \max_{0<j\leq \bar{k}} \left(\prod_{l=1}^{m} ||(J_{m_j}(\lambda_j - \mu_l^{(i)}))^{-1}||_2 \right). \end{aligned} \tag{16}$$

From (13) we know that

$$\min_{1\le l\le k}\min_{1\le j\le m}|\lambda_l - \mu_j^{(i)}| > \delta/2.$$

By using techniques similar to those used to prove (8) in Lemma 3.4, we can bound

$$\begin{aligned}(17)\qquad ||(J_{m_j}(\lambda_j - \mu_l^{(i)}))^{-1}||_2 &\le m_j \max((\delta/2)^{-1}, (\delta/2)^{-m_j})\\ &\le M \max((\delta/2)^{-1}, (\delta/2)^{-\mathrm{M}}),\end{aligned}$$

where $M = \max_{1\le j\le \bar{k}} m_j$. Combining (12), (15), (16), and (17) proves (11). Finally, an induction argument can be used to show that if $\epsilon < 1/(2C)$, $d(\mathcal{S}_i, \mathcal{T}) \le \epsilon$ for all i, $d(\mathcal{S}_{i+1}, \mathcal{T}) < d(\mathcal{S}_i, \mathcal{T})/2$ for all i, and hence $d(\mathcal{S}_i, \mathcal{T}) \to 0$.

If condition (i) or (ii) of Theorem 3.3 is met, then (14) can be replaced by

$$(18)\qquad |\lambda_{k+j} - \mu_{k+j}^{(i)}| \le C_1 ||A_{21}^{(i)}||^2 \le C_1 8\hat{\kappa}_2^2 ||A||_2^2 d(\mathcal{S}_i, \mathcal{T})^2$$

and (15) by

$$(19)\qquad ||p_{i+h+1}(T_2)||_2 \le (3||A||_2)^{m-1} C_1 8\hat{\kappa}_2^2 ||A||_2^2 d(\mathcal{S}_i, \mathcal{T})^2.$$

Combining (12), (19), (16), and (17) proves (4). □

If A_0 has a multiple eigenvalue, there is no guarantee that the m chosen for the degree of the polynomial p_{i+1} is greater than the multiplicity of the eigenvalue. This poses no problem, as is shown by the following theorem.

THEOREM 3.9. *Let $\lambda \in \lambda(A_0)$ have algebraic and geometric multiplicity $m = n - k$. Let $\{A_i\}$ be the sequence generated by Algorithm* 3.1 *using deferred generalized Rayleigh-quotient shifts with polynomials of degree $\bar{m} < m$ and deferring parameter h. Partition the eigenvalues of A_0 into two disjoint sets, $\Lambda_1 = \{\lambda_1, \ldots, \lambda_k\}$ and $\Lambda_2 = \{\lambda, \ldots, \lambda\}$. Define $\mathcal{T}$ to be the invariant subspace corresponding to Λ_1 and let $\mathcal{S}_i$ be the subspace spanned by the first k columns of $\hat{G}_i$. Assume that $\kappa(\hat{G}_i) \le \hat{\kappa}$, some positive constant. Then there exist $\epsilon > 0$ and $C > 0$ so that if $d(\mathcal{S}_i, \mathcal{T}) \le \epsilon, 0 \le i \le h$, then $d(\mathcal{S}_i, \mathcal{T}) \to 0$, and*

$$(20)\qquad d(\mathcal{S}_{i+h+1}, \mathcal{T}) \le C d(\mathcal{S}_{i+h}, \mathcal{T}) d(\mathcal{S}_i, \mathcal{T})^{\bar{m}}.$$

If, in addition, $||A_{12}^{(i)}||_2 = ||A_{21}^{(i)}||_2$ and $A_{22}^{(i)}$ is normal for all i, then

$$(21)\qquad d(\mathcal{S}_{i+h+1}, \mathcal{T}) \le C d(\mathcal{S}_{i+h}, \mathcal{T}) d(\mathcal{S}_i, \mathcal{T})^{2\bar{m}}.$$

Proof. It suffices to show that if ϵ is small enough, $d(\mathcal{S}_i, \mathcal{T}) \le \epsilon$ and $d(\mathcal{S}_{i+h}, \mathcal{T}) \le \epsilon$ implies that

$$(22)\qquad d(\mathcal{S}_{i+h+1}, \mathcal{T}) \le C d(\mathcal{S}_{i+h}, \mathcal{T}) d(\mathcal{S}_i, \mathcal{T})^{\bar{m}}.$$

Again, restrictions on ϵ are indicated throughout the proof.

The proof is identical to the proof of Theorem 3.3, except for the bound on

$$\max_{k<j\le n} ||p_{i+h+1}(\lambda_j)||_2.$$

Partitioning

$$A_i = \begin{pmatrix} \bar{A}_{11}^{(i)} & \bar{A}_{12}^{(i)} \\ \bar{A}_{21}^{(i)} & \bar{A}_{22}^{(i)} \end{pmatrix},$$

where $\bar{A}_{22}^{(i)} \in \mathbf{C}^{\bar{m}\times\bar{m}}$, we can bound

$$\max_{k<j\leq n} ||p_{i+h+1}(\lambda_j)||_2 = ||p_{i+h+1}(\lambda)||_2 \leq \prod_{n-\bar{m}<l\leq n} |\lambda - \mu_l^{(i)}|,$$

where $\lambda(\bar{A}_{11}^{(i)}) = \{\bar{\mu}_j^{(i)}\}_{j=1}^{n-\bar{m}}$ and $\lambda(\bar{A}_{22}^{(i)}) = \{\bar{\mu}_j^{(i)}\}_{j=n-\bar{m}+1}^{n}$. From (9), we know that if ϵ is small enough,

$$||A_{22}^{(i)} - \lambda I||_2 \leq C_1||A_{21}^{(i)}||_2.$$

Since

$$|\lambda - \mu| \leq ||\bar{A}_{22}^{(i)} - \lambda I||_2 \leq ||A_{22}^{(i)} - \lambda I||_2$$

for all $\mu \in \lambda(\bar{A}_{22}^{(i)})$,

$$\prod_{n-\bar{m}<l\leq n} |\lambda - \mu_l^{(i)}| \leq C_1^{\bar{m}}||A_{21}^{(i)}||^{\bar{m}}$$

and (22) follows by carrying this change through in the proof of Theorem 3.3.

The proof when $||A_{12}^{(i)}||_2 = ||A_{21}^{(i)}||_2$ and $A_{22}^{(i)}$ is normal follows similarly. □

4. Convergence rates of deferred shifted methods. Naturally, we would like to compare the convergence rates of the deferred shifted algorithm to the nondeferred shifted algorithm. We concentrate here on the nonsymmetric case.

The Q-convergence rate of a sequence is defined as follows.

DEFINITION 4.1. Let $\{x_k\}$ be a sequence that converges to limit x^*. Then the sequence is said to converge with Q-convergence rate of at least τ if there exists a constant $C > 0$ so that for large k, $|x_{k+1} - x^*| \leq C|x_k - x^*|^\tau$.

Theorem 3.3 shows the Q-convergence rate of the nondeferred shifted GR algorithm to be quadratic ($\tau = 2$), while, if the conditions for (4) are met, the convergence rate is cubic ($\tau = 3$). In general, it is not possible to find the Q-convergence rate of deferred shifted methods. However, there is a related convergence rate that does allow us to perform the comparison.

DEFINITION 4.2. Let $\{x_k\}$ be a sequence that converges to limit x^*. Then the sequence is said to converge with R-convergence rate of at least τ if there exists a sequence that bounds $\{|x_k - x^*|\}$ and converges to zero with Q-convergence rate τ. (This is an alternative definition to the one found in [10].)

To find the R-convergence rate of the deferred shifted algorithm, we need the following well-known result.

THEOREM 4.3. *Let $\{x_k\}$ be a sequence that converges to limit x^*. If there exists k_0 such that for $k \geq k_0$,*

$$||x_{k+1} - x^*||_2 \leq K||x_k - x^*||_2||x_{k-h} - x^*||_2^m, \tag{23}$$

where K and m are constants, then the sequence has an R*-convergence rate of at least τ, where τ is the unique positive root of*

$$t^{h+1} - t^h - m = 0. \tag{24}$$

The proof of this theorem is a generalization of the proof of Theorem 9.2.9 in [10].

An immediate consequence of Theorems 3.3 and 4.3 is now given.

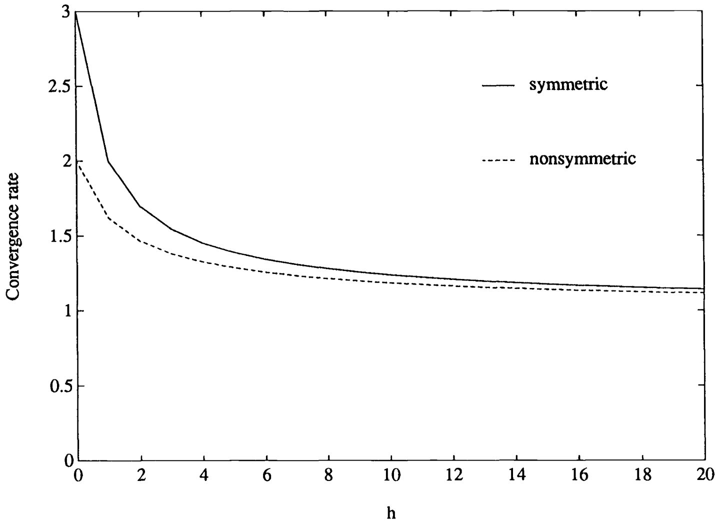


FIG. 1. *Convergence rate of the deferred shifted* GR *method as a function of h.*

COROLLARY 4.4. *Under the conditions in Theorem* 3.3, *the* R*-convergence rate of the* GR *algorithm with deferred shifts equals at least τ_h, the unique positive root of*

$$t^{h+1} - t^h - 1 = 0. \tag{25}$$

If for each iteration, $||A_{12}^{(i)}||_2 = ||A_{21}^{(i)}||_2$ and $A_{22}^{(i)}$ is normal, then the R*-convergence rate equals at least $\bar{\tau}_h$, the unique positive root of*

$$t^{h+1} - t^h - 2 = 0. \tag{26}$$

In particular, $\tau_1 = (1+\sqrt{5})/2$ and $\bar{\tau}_1 = 2$.

The expected deterioration of the convergence rate is plotted in Fig. 1.

As the deferring parameter h increases, we expect to require more iterations before the method in the previous section converges. We now theoretically estimate the amount by which we expect the number of iterations to increase and compare this result with timings from numerical experiments.

The increase in the number of iterations can be estimated as follows: Eventually, the convergence can be modeled by $d(\mathcal{S}_{i+1}, \mathcal{T}) \le C_h d(\mathcal{S}_i, \mathcal{T})^\tau$, where $\tau = \tau_h$. Letting $k = k_h$ be the number of iterations required to obtain the same amount of reduction as one iteration of the nondeferred algorithm, we find that k must satisfy

$$C_h^{(1-\tau^k)/(1-\tau)} d(\mathcal{S}_i, \mathcal{T})^{\tau^k} = C_0 d(\mathcal{S}_i, \mathcal{T})^2,$$

where $\tau = \tau_h$ is the convergence rate of the deferred method. Hence

$$\frac{1-\tau^k}{1-\tau}\log(C_h) + \tau^k \log(d(\mathcal{S}_i, \mathcal{T})) = \log(C_0) + 2\log(d(\mathcal{S}_i, \mathcal{T})).$$

If i is large enough, $\log(d(\mathcal{S}_i, \mathcal{T}))$ will be negative and large in magnitude, and the terms involving C_h and C_0 can be dropped. After canceling $\log(d(\mathcal{S}_i, \mathcal{T}))$, we find

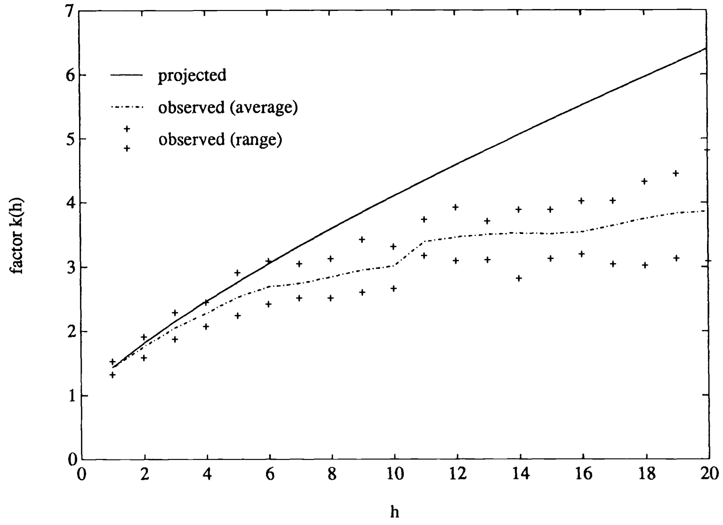


FIG. 2. *Projected factor by which the number of iterations required for convergence of the* GR *algorithm is multiplied vs. ratios of computation time of the deferred and nondeferred double shifted* QR *algorithm, as a function of deferring parameter h.*

that

$$k_h \approx \log(2)/\log(\tau_h).$$

Hence the total number of iterations required can be expected to increase by a factor $\log(2)/\log(\tau_h)$, reported for different h in Fig. 2.

5. Numerical experiments.

To test the effects that deferred shifting has on convergence, we coded a simple double-shifted implicit QR algorithm, which incorporates deflation and deferred shifting with variable deferring parameters, along the lines of Algorithm 7.5.2 in [6]. When this algorithm is executed with $h = 0$, the shift is not deferred, and the execution time is comparable to that of the EISPACK code HQR2. The total computation time was measured for 10 random test matrices of dimension 50 × 50 generated much like the test matrix for the LINPACK benchmark. The matrices were first reduced to upper Hessenberg form, which we did not consider part of the algorithm.

In Table 1, we report the convergence of the $a_{(n-1)(n-2)}$ element for various deferring parameters h for one of the random test matrices. The convergence rates for each iteration are estimated by $\tau_h^{(i)} = \log(a_{(n-1)(n-2)}^{(i)})/\log(a_{(n-1)(n-2)}^{(i-1)})$. They compare favorably with the theoretical results.

In addition, we report the ratios of computation time for the deferred algorithm divided by computation time for the nondeferred algorithms in Fig. 2. It is not surprising that the predicted constant by which the number of iterations required for convergence differs from the measured ratio of the computation times of the deferred versus nondeferred shifted algorithms. After all, while one of the elements of the first subdiagonal converges to zero with convergence rate as predicted in the previous section, the other elements of the first subdiagonal tend to zero as well, although much

TABLE 1

Convergence of element $b^{(i)} = a^{(i)}_{(n-1)(n-2)}$ *of a* 50×50 *random matrix for various deferring parameters* h. *The convergence rate is estimated by* $r_h^{(i)} = \log(a^{(i)}_{(n-1)(n-2)})/\log(a^{(i-1)}_{(n-1)(n-2)})$. *The final row gives the theoretical* R*-convergence rate.*

	$h=0$		$h=1$		$h=5$		$h=10$	
i	$\lvert b^{(i)}\rvert$	$r_0^{(i)}$	$\lvert b^{(i)}\rvert$	$r_1^{(i)}$	$\lvert b^{(i)}\rvert$	$r_5^{(i)}$	$\lvert b^{(i)}\rvert$	$r_{10}^{(i)}$
0	7.1e-01		7.1e-01		7.1e-01		7.1e-01	
1	6.4e-01	1.28	6.4e-01	1.28	6.4e-01	1.28	6.4e-01	1.28
2	5.9e-03	11.60	1.5e-01	4.26	1.5e-01	4.26	1.5e-01	4.26
3	7.0e-05	1.86	3.9e-03	2.94	1.1e-01	1.17	1.1e-01	1.17
4	9.2e-09	1.94	2.6e-04	1.49	3.1e-02	1.57	3.1e-02	1.57
5			7.2e-07	1.71	2.3e-02	1.08	2.3e-02	1.08
6			4.3e-11	1.69	6.1e-03	1.36	6.1e-03	1.36
7					1.0e-04	1.80	4.8e-03	1.05
8					9.8e-06	1.26	1.3e-03	1.25
9					1.3e-07	1.37	1.0e-03	1.03
10					4.8e-09	1.21	2.6e-04	1.20
11							2.1e-04	1.03
12							1.1e-06	1.62
13							2.0e-07	1.13
14							3.4e-09	1.26
r_h		2.00		1.62		1.29		1.18

more slowly. As a result, if the deferred shifted algorithm takes longer to reach the point where the first deflation can occur, the iterate is typically closer to convergence after deflation than the iterate would be after deflation when the nondeferred shifted algorithm is used. This effect is more obvious as the deferring parameter h increases, since as h gets large, the algorithm behaves like an unshifted QR algorithm.

6. Conclusion.

Deferring the shift allows iterations of the algorithm to be pipelined, potentially permitting more efficient use of the processors of a concurrent computer [12], [15]. However, §§4 and 5 show that the convergence is clearly affected by deferring the shift. Whether there is a benefit depends on the specifics of the parallel implementation of the algorithm.

It should be noted that pipelining iterations results in one more complication: the order in which deflation information becomes available will likely require deflation to be delayed just like the shift. This could increase the number of iterations required, since convergence may not be detected until after a new iteration has started. Again, the details of the parallel implementation will determine whether this offsets any gain that resulted from pipelining the iterations.

The main theorem in §3 also shows that the local convergence of nondeferred shifted QR algorithms and their generalizations is quadratic or cubic if the matrix is symmetric or other conditions are met. This remains true even if the matrix has multiple eigenvalues, as long as they are nondefective. This implies that the usual convergence rates can be expected even if the QR algorithm works with a full matrix which has nondefective eigenvalues that are not necessarily simple. In fact, Theorem 3.9 shows that the convergence rate may even improve if an eigenvalue has multiplicity larger than unity.

In addition to the generalized Raleigh-quotient and deferred shifting strategies, the statement of the GR algorithm allows other strategies. After all, $p_{i+1}(x)$ can be chosen however the user pleases. For example, let $c_{i+1}(x)$ represent the characteristic polynomial of A_{22}. One possibility would be to choose $p_{i+1}(x) = c^2_{i+1}(x)$, the square of the characteristic polynomial. A straightforward modification of Theorem 3.3 will show that under the same conditions, the convergence rates for the nondeferred shifted QR algorithm are 3 and 5 for the nonsymmetric and symmetric case, respectively. Naturally, this can be generalized to $p_{i+1}(x) = c_{i+1}^r(x)$, leading to convergence rates of $r + 1$ and $2r + 1$, respectively. In the future, we intend to pursue this strategy, perhaps combined with deferred shifting.

Acknowledgments. The author would like to acknowledge helpful comments from Professors D. P. O'Leary and Alan K. Cline, the author's wife, Dr. Margaret E. Myers, and the referees.

REFERENCES

[1] Z. Bai and J. Demmel, *On a block implementation of Hessenberg multishift* QR *iteration*, LAPACK Working Note 8, ANL, MCS-TM-127, Jan. 1989.

[2] M. Berry and A. Sameh, *Parallel algorithms for the singular value and dense symmetric eigenvalue problems*, CSRD Report No. 761, Univ. of Illinois, Urbana, IL, 1988.

[3] R. Brent and F. Luk, *The solution of singular value and symmetric eigenvalue problems on multi-processor arrays*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 69–84.

[4] J. J. Dongarra and D. C. Sorensen, *A fully parallel algorithm for the symmetric eigenvalue problem*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s139–s154.

[5] P. J. Eberlein, *On the Schur decomposition of a matrix for parallel computation*, Tech. Report, Department of Computer Science, State University of New York, Buffalo, NY, 1985.

[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Second Edition, The Johns Hopkins University Press, Baltimore, MD, 1989.

[7] I. C. F. Ipsen and Y. Saad, *The impact of parallel architectures on the solution of eigenvalue problems*, in Large Scale Eigenvalue Problems, J. Cullum and R. A. Willoughby, eds., Elsevier, Amsterdam, 1986.

[8] S. S. Lo, B. Philippe, and A. Sameh, *A multiprocessor algorithm for the symmetric tridiagonal eigenvalue problem*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s155–s166.

[9] F. T. Luk and H. Park, *On the equivalence and convergence of parallel Jacobi* SVD *algorithms*, in Proc. SPIE, Advanced Algorithms and Architectures for Signal Processing II, Vol. 826, 1988, pp. 152–159.

[10] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations,* Academic Press, New York, 1970.

[11] A. Sameh, *On Jacobi and Jacobi-like algorithms for a parallel computer*, Math. Comp., 25 (1971), pp. 579–590.

[12] G. W. Stewart, *A parallel implementation of the* QR-*algorithm*, Parallel Comput., 5 (1987), pp. 187–196.

[13] ———, *A Jacobi-like algorithm for computing the Schur decomposition of a non-Hermitian matrix*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 853–864.

[14] ———, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[15] R. A. van de Geijn, *Implementing the* QR-*algorithm on an array of processors*, Ph.D. thesis, TR-1897, Department of Computer Science, University of Maryland, College Park, MD, 1987.

[16] R. A. van de Geijn and D. G. Hudson, *An efficient parallel implementation of the nonsymmetric* QR *algorithm,* in Proc. Fourth Conference on Hypercube Concurrent Computers and Applications, Golden Gate Enterprises, Los Altos, CA, 1990.

[17] D. S. Watkins and L. Elsner, *Convergence of algorithms of decomposition type for the eigenvalue problem*, Linear Algebra Appl., 143 (1991), pp. 19–47.

[18] ———, *Chasing algorithms for the eigenvalue problem*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 374–384.

[19] J. H. Wilkinson, *The Algebraic Eigenvalue Problem,* Oxford University Press, London, U.K., 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 195–234, January 1993

017

BOUNDS ON PERTURBATIONS OF GENERALIZED SINGULAR VALUES AND OF ASSOCIATED SUBSPACES*

REN-CANG LI[†]

Abstract. The sensitivity of the generalized singular value decomposition of a matrix pair to perturbations in the matrix elements is analysed. It is shown how the chordal distances between the singular values and the angular distance between the generalized singular spaces can be bounded in terms of the angular distances between the matrices. The main results are generalizations of several results on the standard singular value problem.

Key words. Grassmann matrix pair, generalized singular value decomposition, generalized singular subspace, the chordal metric, metrics on Grassmann manifold, unitarily invariant norm, perturbation bound

AMS(MOS) subject classifications. 15A18, 15A22, 15A23, 15A42

1. Introduction.

The generalized singular value decomposition (GSVD) for a matrix pair of two matrices with the same number of columns was proposed by Van Loan in 1976. This article addresses the following question: when a matrix pair is perturbed, by how much can its generalized singular values (GSVs) and subspaces associated with its GSVD change? This problem was first analysed by Sun in 1983. (Paige also gave a bound for GSV variations in 1984.) Our main results that are different from those of Sun and Paige are generalizations of several results on the standard singular value problem.

Throughout the paper, capital letters are for matrices, lowercase Latin letters are for column vectors or scalars, and lowercase Greek letters are for scalars; $\mathbb{C}^{m\times n}$ is the set of $m \times n$ complex matrices; $\mathcal{U}_n \subset \mathbb{C}^{n\times n}$ is the set of $n \times n$ unitary matrices, $\mathbb{C}^m = \mathbb{C}^{m\times 1}$, $\mathbb{C} = \mathbb{C}^1$; $\mathbb{R}$ is the real number set. The symbol $I^{(n)}$ stands for the $n \times n$ unit matrix, and $0_{m,n}$ for the $m \times n$ null matrix (also we just write I and 0 for convenience when no confusion arises). A^T, A^H, and A^+ denote the transpose, conjugate transpose, and Moore–Penrose inverse of A, respectively. $\mathcal{R}(X)$ is the column space, the subspace spanned by the column vectors of X, and P_X is the orthogonal projection onto the column space $\mathcal{R}(X)$. It is easy to verify that

$$P_X = XX^+, \qquad P_{X^H} = X^+X.$$

We will consider *unitarily invariant norms* $\|\cdot\|$ of matrices. In this we follow [34, pp. 74–87]. To say that the norm is *unitarily invariant* on $\mathbb{C}^{m\times n}$ means it satisfies, besides the usual properties of any norm,

(1) $\|UAV\| = \|A\|$, for any $U \in \mathcal{U}_m$, and $V \in \mathcal{U}_n$.

(2) $\|A\| = \|A\|_2$, for any $A \in \mathbb{C}^{m\times n}$, rank $A = 1$.

Two unitarily invariant norms used frequently are the spectral norm $\|\cdot\|_2$ and the Frobenius norm $\|\cdot\|_F$.

*Received by the editors October 30, 1989; accepted for publication (in revised form) July 31, 1991.

[†]Institute of Applied Physics and Computational Mathematics, P. O. Box 8009, Beijing 100088, People's Republic of China. Present address, Department of Mathematics, University of California, Berkeley, California 94720 (li@math.berkeley.edu).

In this paper, very often matrices with different dimensions enter our arguments together, so we adhere to the agreement made on [34, p. 79]. In this way, we have

$$\|CD\| \le \begin{cases} \|C\|_2\|D\| \\ \|C\|\|D\|_2 \end{cases}, \quad \text{for any } C \in \mathbb{C}^{m\times n},\ D \in \mathbb{C}^{n\times \ell}. \tag{1.1}$$

This inequality will be used extensively below. Henceforth, the symbol $\|\cdot\|$ without any subscripts is reserved for a unitarily invariant norm.

For any $A \in \mathbb{C}^{m\times n}$, its *singular value decomposition* (SVD) may be written as (see, e.g., [34] and [7])

$$A = U\Sigma V^H, \qquad \Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \dots), \tag{1.2}$$

where $U \in \mathcal{U}_m$, $V \in \mathcal{U}_n$, and $\sigma_1 \ge \cdots \ge 0$ are singular values (SVs) of A. With the help of SVD, any unitarily invariant norm can be written as $\|A\| = \Phi(\sigma_1, \sigma_2, \dots)$ where Φ is a *symmetric gauge function* (see, e.g., [23] and [34]). The set of SVs of A is denoted by $\sigma(A)$. Due to the numerical stability of SVD, it has been applied to a variety of practical problems with moderate dimensions. It is an especially major tool for solving ill-conditioned problems in linear algebra. There are quite a number of excellent papers written for various topics relating to SVD and its applications in the literature; see, e.g., [7], [8], [9], [22], [23], [27], [34], [36], [40], and other references therein.

Motivated by SVD, Van Loan [38] and Paige and Saunders [25] suggested several forms of the GSVD of two matrices having the same number of columns. GSVD immediately attracted the attention of a number of numerical analysts. Now, a few algorithms for its numerical computation are available, and applications to practical problems are also being made; see, e.g., [10], [12], [13], [14], and [38].

As to the perturbation theory for the standard singular value problem, the following result has been known for nearly thirty years. It was proved by Mirsky [23] with the help of a powerful theorem for Hermitian matrices of Lidskii [21] and Wielandt [42]. Let A, $\widetilde{A} \in \mathbb{C}^{m\times n}$, and let $\alpha_1 \ge \cdots \ge \alpha_q$ and $\widetilde{\alpha}_1 \ge \cdots \ge \widetilde{\alpha}_q$ ($q = \min\{m, n\}$) be their SVs, respectively. Then for any unitarily invariant norm $\|\cdot\|$

$$\left\|\operatorname{diag}(\alpha_1 - \widetilde{\alpha}_1, \dots, \alpha_q - \widetilde{\alpha}_q)\right\| \le \|A - \widetilde{A}\|. \tag{1.3}$$

Also, Wedin [40] has generalized the celebrated Davis–Kahan $\sin\theta$ theorems [4] relating to the invariant subspaces of two self-adjoint operators to cover the SVD. A problem which naturally arises is how the generalized singular values (GSVs) and associated subspaces behave under a perturbation. In 1983, Sun [30] gave a detailed analysis of this problem. In that paper, he generalized several noted results for the standard singular value problem. Although [30] is an excellent paper, there are still a few questions left to be answered. For example, the generalization of (1.3) to cover a general unitarily invariant norm is still open. Part I of this paper is written for this purpose.

The paper is divided into two parts. In the first part, we concentrate our attention on the perturbations of GSVs, and in the second part we focus on the perturbations of subspaces associated with GSVD. The main idea comes from Li's recent papers [17], [18]. Our results are completely different from those in Sun [30].

Now, we outline definitions relating to GSVD. Readers are referred to Sun [30] for the motivations of these definitions.

DEFINITION 1.1. Let A, $B \in \mathbb{C}^{n\times n}$. $A - \lambda B$ is called a *regular matrix pencil of order* n if $\det(A - \lambda B) \not\equiv 0$ for $\lambda \in \mathbb{C}$. A complex number pair $(\alpha, \beta) \neq (0,0)$ is termed a *generalized eigenvalue* of a regular matrix pencil $A - \lambda B$ if $\det(\beta A - \alpha B) = 0$. Denote by $\lambda(A, B)$ the set of the generalized eigenvalues (counted according to their algebraic multiplicities) of $A - \lambda B$.

DEFINITION 1.2. Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times n}$. A matrix pair $\{A, B\}$ is an (m, p, n)-*Grassmann matrix pair* (GMP) if $\operatorname{rank}\binom{A}{B} = n$.

DEFINITION 1.3. Let $\{A, B\}$ be an (m, n, p)-GMP. A nonnegative number-pair (α, β) is a GSV of the GMP $\{A, B\}$ if

$$(1.4a) \qquad (\alpha, \beta) \neq (0,0), \quad \det(\beta^2 A^H A - \alpha^2 B^H B) = 0, \quad \alpha,\ \beta \geq 0,$$

i.e.,

$$(1.4b) \qquad (\alpha, \beta) = (\sqrt{\lambda}, \sqrt{\mu}), \quad \text{where } (\lambda, \mu) \in \lambda(A^H A, B^H B) \quad \text{and} \quad \lambda,\ \mu \geq 0.$$

The set of GSV of $\{A, B\}$ is denoted by $\sigma\{A, B\}$.

The fact we should bear in mind in giving (1.4b) is that for H, $K \in \mathbb{C}^{n\times n}$, H, $K \geq 0$, if $(\lambda, \mu) \in \lambda(H, K)$, then $(|\lambda|, |\mu|) \in \lambda(H, K)$.

Clearly, if $\{A, B\}$ is an (m, p, n)-GMP, then $A^H A - \lambda B^H B$ is a definite matrix pencil, i.e., $x^H A^H A x + x^H B^H B x > 0$ for all $0 \neq x \in \mathbb{C}^n$, and vice versa. The definite pencil $A^H A - \lambda B^H B$ has n generalized eigenvalues, hence GMP $\{A, B\}$ has n GSVs. A well-developed perturbation theory for the generalized eigenvalue problem of definite pencils is available; see, e.g., Sun [31], [32]; Stewart [28]; and Li [17]. Perturbation bounds for the generalized singular value problem can be obtained with the help of the close relation between the two problems. However, the bounds obtained in this way are often unsatisfactory, just like the perturbation bounds for the SVs of a single matrix A obtained through the perturbation bounds for the eigenvalues of the Hermitian matrix $A^H A$. So special attention deserves to be paid to perturbations for the generalized singular value problem.

GSVD has several forms in the literature. In this paper, we adopt the following form (see Van Loan [38], Paige and Saunders [25], and Sun [30]).

THEOREM 1.1 (GSVD). *Let $\{A, B\}$ be an (m, p, n)-GMP. Then there exist matrices $U \in \mathcal{U}_m$, $V \in \mathcal{U}_p$, and $Q \in \mathbb{C}^{n\times n}$ nonsingular such that*

$$(1.5a) \qquad U^H A Q = \Sigma_A, \qquad V^H B Q = \Sigma_B,$$

where

$$(1.5b) \qquad \Sigma_A = \operatorname{diag}(\alpha_1, \alpha_2, \ldots),$$

$$(1.5c) \qquad \Sigma_B = \operatorname{diag}(\ldots, \beta_{n-1}, \beta_n),$$

meaning that Σ_A is zero except for the diagonal starting in the top left corner (*leading diagonal*), *and Σ_B is zero except for the diagonal finishing in the bottom right corner* (*trailing diagonal*). *$\alpha_{m+1} = \cdots = \alpha_n = 0$, if $m \leq n$; $\beta_1 = \cdots = \beta_{n-p} = 0$, if $p \leq n$; and α_i, $\beta_i \geq 0$, $\alpha_i^2 + \beta_i^2 = 1$, $i = 1, 2, \ldots, n$.*

By Definition 1.3, we have $\sigma\{A, B\} = \{(\alpha_i, \beta_i),\ i = 1, 2, \ldots, n\}$ for the (m, p, n)-GMP $\{A, B\}$ in Theorem 1.1.

LEMMA 1.1. *If $\{A, B\}$ is as described in Theorem 1.1, then for any unitarily invariant norm $\|\cdot\|$,*

$$\|Q\| \leq \|Z^+\| \quad \textit{and} \quad \|Q^{-1}\| \leq \|Z\|, \tag{1.6}$$

where $Z = \binom{A}{B}$.

Proof. Bearing $\operatorname{rank} Z = n$ and $Z^+Z = I$ in mind, we have

$$\begin{pmatrix} U^H & \\ & V^H \end{pmatrix} ZQ = \begin{pmatrix} \Sigma_A \\ \Sigma_B \end{pmatrix} \Rightarrow \begin{cases} Q = Z^+ \begin{pmatrix} U & \\ & V \end{pmatrix} \begin{pmatrix} \Sigma_A \\ \Sigma_B \end{pmatrix}, \\ Q^{-1} = (\Sigma_A^H, \Sigma_B^H) \begin{pmatrix} U^H & \\ & V^H \end{pmatrix} Z. \end{cases}$$

Now, from the above equations, (1.1), and

$$\left\| \begin{pmatrix} \Sigma_A \\ \Sigma_B \end{pmatrix} \right\|_2 = \|(\Sigma_A^H, \Sigma_B^H)\|_2 = 1,$$

(1.6) follows. □

Remark 1.1. We note that in order to have rank n, $m + p \geq n$. When $m + p = n$ we can choose $Q = Z^{-1}$ and get a trivial GSVD with

$$\sigma\{A, B\} = \{\underbrace{(1,0), \ldots, (1,0)}_{m}, \underbrace{(0,1), \ldots, (0,1)}_{p=n-m}\}. \tag{1.7}$$

We use *the chordal metric on the Riemann sphere* to measure the difference between two generalized singular values, and *metrics on the Grassmann manifold* to measure the difference between two matrix pairs (see Sun [29]–[35] and Li [16]–[18]). For $(\alpha, \beta) \neq (0,0)$, $(\gamma, \delta) \neq (0,0)$, *the chordal distance* between the two points is defined by

$$\rho\big((\alpha, \beta), (\gamma, \delta)\big) \overset{\mathrm{def}}{=} \frac{|\delta\alpha - \gamma\beta|}{\sqrt{|\alpha|^2 + |\beta|^2}\sqrt{|\gamma|^2 + |\delta|^2}}. \tag{1.8}$$

Let $X, Y \in \mathbb{C}^{m \times n}$ $(m > n)$ both have full column rank n, and define the angle $\Theta(X, Y)$ between X and Y as [34]

$$\Theta(X, Y) \overset{\mathrm{def}}{=} \arccos((X^HX)^{-\frac{1}{2}}X^HY(Y^HY)^{-1}Y^HX(X^HX)^{-\frac{1}{2}})^{-\frac{1}{2}} \geq 0. \tag{1.9}$$

Here and in the following, $A > 0$ $(A \geq 0)$ denotes that A is a positive definite (nonnegative definite) Hermitian matrix. $A^{1/2}$ is the unique positive definite (nonnegative definite) square root of $A \geq 0$; and $A^{-1/2} = (A^{1/2})^{-1}$ for $A > 0$. The difference between these two points in the Grassmann manifold can be measured by appropriate unitarily invariant norms of $\sin\Theta(X, Y)$ or of $P_X - P_Y$. The reader is referred to Lemma 3.1 below for the relations between the unitarily invariant norms of $\sin\Theta(X, Y)$ or those of $P_X - P_Y$.

Lemma 1.2. *Let $X, Y \in \mathbb{C}^{m\times n}$ $(1 \le n \le m-1)$ have full column rank n. Suppose that $\widetilde{Y} = (Y, Y_1) \in \mathbb{C}^{m\times m}$ is a nonsingular matrix with*

$$\widetilde{Y}^{-1} = \begin{pmatrix} S^H \\ S_1^H \end{pmatrix}, \qquad S \in \mathbb{C}^{m\times n}.$$

Then for any unitarily invariant norm $\|\cdot\|$,

$$\|\sin\Theta(X,Y)\| = \|(S_1^H S_1)^{-\frac{1}{2}} S_1^H X (X^H X)^{-\frac{1}{2}}\|.$$

Lemma 1.2 is now well known. For a proof of it, the reader is referred to, e.g., Li [17, Lemma 2.1].

Throughout this paper, $\{A, B\}$ and $\{\widetilde{A}, \widetilde{B}\}$ are always reserved for two (m,p,n)-GMPs except when otherwise stated.

Part I. Perturbation bounds for GSV. In §2 we study a geometric representation of a GSV and pairing problems for two sets of GSVs. The former is similar to the statements in [17] for a real generalized eigenvalue, and the latter may be of general interest apart from the role it plays in this paper. We present the main results concerning GSV in §3 and give their proofs in §4.

2. Geometric representation and pairing. By Definition 1.3, it is evident that every GSV can be represented by a pair (α, β) of real numbers α and β satisfying

$$\alpha, \beta \ge 0, \qquad \alpha^2 + \beta^2 = 1. \tag{2.1}$$

Thus a 1 to 1 correspondence between the set of GSVs and a quarter-circular arc Γ of the unit circle is established in the following way. (α,β) satisfying (2.1) corresponds to a point $z \in \Gamma$, as shown in Fig. 1. Therefore, every $z \in \Gamma$ determines a number pair (α, β) satisfying (2.1) by its coordinate. Moreover, (α,β) determines a unique angle $\theta = \theta(\alpha,\beta) = \arccos\alpha = \arcsin\beta$, $(0 \le \theta \le \frac{\pi}{2})$. For convenience, in the following text, we treat z and (α,β) equally and write $z = (\alpha,\beta)$. The symbol Γ is always reserved for the quarter-circular arc.

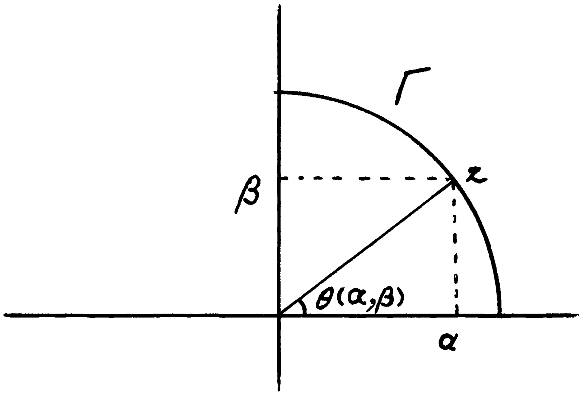


Fig. 1. *Geometric representation: $z = (\alpha, \beta)$.*

Given (α,β), (γ,δ) corresponding to z, $w \in \Gamma$, the notation

$$(\alpha,\beta) \prec (\gamma,\delta) \qquad ((\alpha,\beta) \preceq (\gamma,\delta)) \tag{2.2}$$

means

$$\theta(\alpha,\beta) < \theta(\gamma,\delta) \qquad (\theta(\alpha,\beta) \leq \theta(\gamma,\delta)); \tag{2.3}$$

and the notation $z \prec w$ ($z \preceq w$) means that (2.2) holds. We will use the notation (zwv) to mean that the points z, w, $v \in \Gamma$ appear in the order $z \preceq w \preceq v$. Similarly, we can deal with more than three points. A distance on Γ is defined by (refer to (1.8))

$$d(z,w) \stackrel{\text{def}}{=} \rho\big((\alpha,\beta),(\gamma,\delta) \big).$$

It is easy to prove that (Γ, d) is a complete distance space (see Proposition 2.1 below).

PROPOSITION 2.1 (Li [17]). *Let (α,β), (γ,δ) be two* GSVs. *Then*

$$\rho\big((\alpha,\beta),(\gamma,\delta) \big) = \sin|\theta(\alpha,\beta) - \theta(\gamma,\delta)|.$$

Now, we consider the pairing problem for two sets of GSVs.

PROPOSITION 2.2. *Let (α_i,β_i) and (γ_j,δ_j), $i,j = 1,2,\dots,n$ be $2n$* GSV*s arranged in increasing order, respectively, i.e.,*

$$(\alpha_1,\beta_1) \preceq \cdots \preceq (\alpha_n,\beta_n) \quad \textit{and} \quad (\gamma_1,\delta_1) \preceq \cdots \preceq (\gamma_n,\delta_n). \tag{2.4}$$

Then

$$\min_{\tau} \max_{1\leq j\leq n} \rho\big((\alpha_j,\beta_j),(\gamma_{\tau(j)},\delta_{\tau(j)}) \big) = \max_{1\leq j\leq n} \rho\big((\alpha_j,\beta_j),(\gamma_j,\delta_j) \big). \tag{2.5}$$

The minimum in (2.5) *is taken over all permutations of* $\{1,2,\dots,n\}$.

Proposition 2.2 may be of great importance. The reader will find that all our bounds in §3 involve permutations of $\{1,2,\dots,n\}$. Generally, except for their existence, they are unknown. Proposition 2.2 may be the only result now available which tells us what the exact permutations are in a few cases (refer to Theorems 3.1 and 3.2, Corollary 3.1, Remark 3.1, and Lemma 4.2). An insight into the importance of Proposition 2.2 can also be learned by noting some facts arising from the perturbation theories for the standard eigenvalue and singular value problems. Recall that for $2n$ real numbers $\alpha_1 \leq \cdots \leq \alpha_n$ and $\gamma_1 \leq \cdots \leq \gamma_n$, we have

$$\max_{1\leq j\leq n} |\alpha_j - \gamma_j| = \min_{\mu} \max_{1\leq j\leq n} |\alpha_j - \gamma_{\mu(j)}|,$$

and more generally, for any unitarily invariant norm $\|\cdot\|$,

$$\big\|\operatorname{diag}(\alpha_1 - \gamma_1,\dots,\alpha_n - \gamma_n)\big\| = \min_{\mu}\big\|\operatorname{diag}(\alpha_1 - \gamma_{\mu(1)},\dots,\alpha_n - \gamma_{\mu(n)})\big\|,$$

where $\min_\mu$ is taken over all permutations of $\{1,2,\dots,n\}$. If α_i and γ_j are no longer real but may be complex, the situation becomes very complicated, and the above two equations are no longer true. The fact that we have not yet completely solved an open problem [23] of extending the well-known Weyl–Lidskii theorem (see, e.g., [34]) to the spectral perturbation of a normal matrix should be partially attributed to the fact that we do not know what a proper pairing of $2n$ complex numbers is. (Much work on such extensions has been done by several mathematicians; see, e.g., [1] and [2].)

The following proposition will help us to finish our proofs for the nonsquare case in §4.

PROPOSITION 2.3. *Let* (α_i, β_i) *and* (γ_j, δ_j), $i, j = 1, 2, \ldots, n$, *be* $2n$ GSVs; *let* p, $q \geq 0$ *be two integers; and define*

$$\begin{aligned}(\alpha_{n+i}, \beta_{n+i}) = (\gamma_{n+i}, \delta_{n+i}) = (1, 0), \qquad i = 1, 2, \ldots, p;\\ (\alpha_{n+p+j}, \beta_{n+p+j}) = (\gamma_{n+p+j}, \delta_{n+p+j}) = (0, 1), \qquad j = 1, 2, \ldots, q.\end{aligned}$$

Also, let μ *be a permutation of* $\{1, 2, \ldots, n+p+q\}$. *Then there exists a permutation* τ *of* $\{1, 2, \ldots, n\}$ *such that for any unitarily invariant norm* $\|\cdot\|$

$$\begin{aligned}&\left\|\mathrm{diag}\,\Big(\rho\big((\alpha_1, \beta_1), (\gamma_{\mu(1)}, \delta_{\mu(1)})\big), \ldots, \rho\big((\alpha_{n+p+q}, \beta_{n+p+q}), (\gamma_{\mu(n+p+q)}, \delta_{\mu(n+p+q)})\big)\Big)\right\|\\ &\qquad \geq \left\|\mathrm{diag}\,\Big(\rho\big((\alpha_1, \beta_1), (\gamma_{\tau(1)}, \delta_{\tau(1)})\big), \ldots, \rho\big((\alpha_n, \beta_n), (\gamma_{\tau(n)}, \delta_{\tau(n)})\big)\Big)\right\|.\end{aligned} \tag{2.6}$$

In order to keep the paper fairly short, we do not give the details of the proofs of Propositions 2.2 and 2.3, which are based on constructive arguments.

Proposition 2.2 solves the problem: Find a permutation τ of $\{1, 2, \ldots, n\}$ such that

$$\begin{aligned}&\min_{\mu} \max_{1\leq j\leq n} \rho\big((\alpha_j, \beta_j), (\gamma_{\mu(j)}, \delta_{\mu(j)})\big)\\ &\quad \equiv \min_{\mu} \left\|\mathrm{diag}\,\Big(\rho\big((\alpha_1, \beta_1), (\gamma_{\mu(1)}, \delta_{\mu(1)})\big), \ldots, \rho\big((\alpha_n, \beta_n), (\gamma_{\mu(n)}, \delta_{\mu(n)})\big)\Big)\right\|_2\\ &\quad = \left\|\mathrm{diag}\,\Big(\rho\big((\alpha_1, \beta_1), (\gamma_{\tau(1)}, \delta_{\tau(1)})\big), \ldots, \rho\big((\alpha_n, \beta_n), (\gamma_{\tau(n)}, \delta_{\tau(n)})\big)\Big)\right\|_2\\ &\quad \equiv \max_{1\leq j\leq n} \rho\big((\alpha_j, \beta_j), (\gamma_{\tau(j)}, \delta_{\tau(j)})\big).\end{aligned} \tag{2.7}$$

Let ς and ν be two permutations of $\{1, 2, \ldots, n\}$ with the property

$$(\alpha_{\varsigma(1)}, \beta_{\varsigma(1)}) \preceq \cdots \preceq (\alpha_{\varsigma(n)}, \beta_{\varsigma(n)}) \quad \text{and} \quad (\gamma_{\nu(1)}, \delta_{\nu(1)}) \preceq \cdots \preceq (\gamma_{\nu(n)}, \delta_{\nu(n)});$$

then $\tau = \nu\varsigma^{-1}$ is a solution to the problem. In other words, τ satisfies (2.7). Generally, there is no global solution (independent of $\|\cdot\|$) to the optimization problem

$$\min_{\mu} \left\|\mathrm{diag}\,\Big(\rho\big((\alpha_1, \beta_1), (\gamma_{\mu(1)}, \delta_{\mu(1)})\big), \ldots, \rho\big((\alpha_n, \beta_n), (\gamma_{\mu(n)}, \delta_{\mu(n)})\big)\Big)\right\| \tag{2.8}$$

if the minimum is taken over all permutations of $\{1, 2, \ldots, n\}$. To see this, we give a counterexample: $n = 2$, $z_1 = (\alpha_1, \beta_1) = (1, 0)$, $z_2 = (\alpha_2, \beta_2) = (\sqrt{3}/2, 1/2)$, $w_1 = (\gamma_1, \delta_1) = (1/2, \sqrt{3}/2)$, $w_2 = (\gamma_2, \delta_2) = (0, 1)$, and $\|\cdot\| = \|\cdot\|_F$. It is easy to verify that $z_1 \prec z_2$, $w_1 \prec w_2$. Therefore ς, ν and $\tau = \nu\varsigma^{-1}$ are all the identity permutation of $\{1, 2\}$. Only two permutations exist: one is the identity permutation ω, i.e., $\omega(i) = i$ for $i = 1, 2$; the other is μ defined by $\mu(1) = 2$, $\mu(2) = 1$. It is easy to verify that

$$\begin{aligned}&\left\|\mathrm{diag}\,\Big(\rho\big((\alpha_1, \beta_1), (\gamma_{\omega(1)}, \delta_{\omega(1)})\big), \rho\big((\alpha_2, \beta_2), (\gamma_{\omega(2)}, \delta_{\omega(2)})\big)\Big)\right\|_2\\ &\quad = \frac{\sqrt{3}}{2} < 1 = \left\|\mathrm{diag}\,\Big(\rho\big((\alpha_1, \beta_1), (\gamma_{\mu(1)}, \delta_{\mu(1)})\big), \rho\big((\alpha_2, \beta_2), (\gamma_{\mu(2)}, \delta_{\mu(2)})\big)\Big)\right\|_2,\end{aligned}$$

whereas

$$\left\|\text{diag}\,\big(\rho\big(\,(\alpha_1,\beta_1),(\gamma_{\omega(1)},\delta_{\omega(1)})\big),\rho\big(\,(\alpha_2,\beta_2),(\gamma_{\omega(2)},\delta_{\omega(2)})\big)\big)\right\|_F$$
$$=\frac{\sqrt{6}}{2}>\frac{\sqrt{5}}{2}=\left\|\text{diag}\,\big(\rho\big(\,(\alpha_1,\beta_1),(\gamma_{\mu(1)},\delta_{\mu(1)})\big),\rho\big(\,(\alpha_2,\beta_2),(\gamma_{\mu(2)},\delta_{\mu(2)})\big)\big)\right\|_F.$$

Remark 2.1. $\|\cdot\|_2$ is unitarily invariant, and for $\|\cdot\|_2$, Proposition 2.3 is a corollary of Proposition 2.2. To see this, we note that the spectral norm of a diagonal matrix is the maximal modulus of its diagonal elements (refer to (2.7)), and that the values of the qualities before $\geq$ in the inequality (2.6) are reduced if the permutations μ are replaced by those optimal ones obtained analogously to the permutation τ of Proposition 2.2 by arranging (α_i,β_i) and (γ_j,δ_j) in ascending order. Under these optimal permutations, $(1,0)$ must be paired to $(1,0)$, and $(0,1)$ to $(0,1)$; and thus $(1,0)$ and $(0,1)$ can be eliminated as required. We note that for a general unitarily invariant norm, Proposition 2.3 may not be regarded as a corollary of Proposition 2.2 (refer to the counterexample above).

Remark 2.2. Proposition 2.3 plays an important role in our proofs in §4. In fact, in order to complete our proof for the cases $m\neq n$ and/or $p\neq n$, we augment A and B of an (m,p,n)-GMP $\{A,B\}$ suitably to a bigger GMP $\{\widehat{A},\widehat{B}\}$, with $\widehat{A}$, $\widehat{B}$ being square, and $\sigma\{\widehat{A},\widehat{B}\}$ being a union of $\sigma\{A,B\}$ with a few $(1,0)$ and/or $(0,1)$. Proposition 2.3 guarantees the possibility of removing the added $(1,0)$ and/or $(0,1)$ from the bounds via those for the case $m=p=n$ (refer to §4).

3. Main perturbation theorems for GSV.

THEOREM 3.1. *Let* $\{A,B\}$ *and* $\{\widetilde{A},\widetilde{B}\}$ *be two* (m,p,n)-GMP*s, and let* $\sigma\{A,B\}=\{(\alpha_i,\beta_i),\,i=1,2,\cdots,n\}$ *and* $\sigma\{\widetilde{A},\widetilde{B}\}=\{(\widetilde{\alpha}_j,\widetilde{\beta}_j),\,j=1,2,\ldots,n\}$. *Then there exists a permutation* τ *of* $\{1,2,\ldots,n\}$ *such that*

$$\max_{1\leq j\leq n}\rho\big(\,(\alpha_j,\beta_j),(\widetilde{\alpha}_{\tau(j)},\widetilde{\beta}_{\tau(j)})\big)\leq\|\sin\Theta(Z,\widetilde{Z})\|_2, \tag{3.1}$$

where

$$Z=\begin{pmatrix}A\\B\end{pmatrix},\qquad \widetilde{Z}=\begin{pmatrix}\widetilde{A}\\\widetilde{B}\end{pmatrix}\in\mathbb{C}^{(m+p)\times n}. \tag{3.2}$$

THEOREM 3.2. *The conditions and notations are as described in Theorem* 3.1. *Then there exists a permutation* μ *of* $\{1,2,\ldots,n\}$ *such that*

$$\sqrt{\sum_{j=1}^{n}\Big[\rho\big(\,(\alpha_j,\beta_j),(\widetilde{\alpha}_{\mu(j)},\widetilde{\beta}_{\mu(j)})\big)\Big]^2}\leq\|\sin\Theta(Z,\widetilde{Z})\|_F. \tag{3.3}$$

For two square matrix pairs, we have Theorem 3.3.

THEOREM 3.3. *To the assumptions of Theorem* 3.1, *we add that* $m=p=n$. *Then there exists a permutation* τ *of* $\{1,2,\ldots,n\}$ *such that*

$$\max_{1\leq j\leq n}\rho\big(\,(\alpha_j,\beta_j),(\widetilde{\alpha}_{\tau(j)},\widetilde{\beta}_{\tau(j)})\big)$$
$$=\min_{U,V\in\mathcal{U}_n}\left\|(Z^HZ)^{-\frac{1}{2}}(A^HV\widetilde{B}-B^HU\widetilde{A})(\widetilde{Z}^H\widetilde{Z})^{-\frac{1}{2}}\right\|_2. \tag{3.4}$$

We present two corollaries obtained from Theorems 3.1 and 3.2. But first we need the following lemma.

LEMMA 3.1. *If $X, Y \in \mathbb{C}^{q\times t}$ $(q > t)$ have full column rank t, then $P_X^{\perp}(Y - X)(Y^HY)^{-\frac{1}{2}}$, $P_Y^{\perp}(Y-X)(X^HX)^{-\frac{1}{2}}$, and $\sin\Theta(X,Y)$ have the same nonzero singular values, where $P_X^{\perp} = I - P_X$, and $P_Y^{\perp} = I - P_Y$. Moreover if their nonzero singular values are $\sigma_1, \sigma_2, \ldots$, the nonzero singular values of $P_Y - P_X$ are $\sigma_1, \sigma_1, \sigma_2, \sigma_2, \ldots$. Therefore,*

$$\begin{aligned}\|P_Y - P_X\|_2 &= \|P_X^{\perp}(Y - X)(Y^HY)^{-\frac{1}{2}}\|_2 = \|P_Y^{\perp}(Y - X)(X^HX)^{-\frac{1}{2}}\|_2 \\ &= \|\sin\Theta(X,Y)\|_2,\end{aligned} \tag{3.5a}$$

(3.5b)

$$\begin{aligned}\frac{1}{\sqrt{2}}\|P_Y - P_X\|_F &= \|P_X^{\perp}(Y - X)(Y^HY)^{-\frac{1}{2}}\|_F = \|P_Y^{\perp}(Y - X)(X^HX)^{-\frac{1}{2}}\|_F \\ &= \|\sin\Theta(X,Y)\|_F.\end{aligned}$$

Proof. Choose $U \in \mathcal{U}_q$ such that

$$U^HX = \begin{pmatrix} X_1 \\ 0 \end{pmatrix}, \quad U^HY = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}, \quad X_1, Y_1 \in \mathbb{C}^{t\times t}.$$

Obviously, X_1 is nonsingular. Thus we have

$$\begin{aligned}P_Y - P_X &= Y(Y^HY)^{-1}Y^H - X(X^HX)^{-1}X^H \\ &= U\begin{pmatrix} Y_1(Y_1^HY_1 + Y_2^HY_2)^{-1}Y_1^H - I^{(t)} & Y_1(Y_1^HY_1 + Y_2^HY_2)^{-1}Y_2^H \\ Y_2(Y_1^HY_1 + Y_2^HY_2)^{-1}Y_1^H & Y_2(Y_1^HY_1 + Y_2^HY_2)^{-1}Y_2^H \end{pmatrix}U^H\end{aligned}$$

and

$$\begin{aligned}&(P_Y - P_X)^H(P_Y - P_X) = (P_Y - P_X)^2 \\ &\quad = U\begin{pmatrix} I^{(t)} - Y_1(Y_1^HY_1 + Y_2^HY_2)^{-1}Y_1^H & 0 \\ 0 & Y_2(Y_1^HY_1 + Y_2^HY_2)^{-1}Y_2^H \end{pmatrix}U^H.\end{aligned} \tag{3.6}$$

On the other hand, from

$$P_X^{\perp} = U\begin{pmatrix} 0 & 0 \\ 0 & I^{(q-t)} \end{pmatrix}U^H$$

follows

$$\begin{aligned}&\big(P_X^{\perp}(Y - X)\big)(Y^HY)^{-1}\big(P_X^{\perp}(Y - X)\big)^H \\ &\qquad = U\begin{pmatrix} 0 & 0 \\ 0 & Y_2(Y_1^HY_1 + Y_2^HY_2)^{-1}Y_2^H \end{pmatrix}U^H.\end{aligned} \tag{3.7}$$

Note also that

$$\begin{aligned}\underline{\lambda}(Y_2(Y_1^HY_1 + Y_2^HY_2)^{-1}Y_2^H) &= \underline{\lambda}(Y_2^HY_2(Y_1^HY_1 + Y_2^HY_2)^{-1}) \\ &= \underline{\lambda}(I - Y_1^HY_1(Y_1^HY_1 + Y_2^HY_2)^{-1}) \\ &= \underline{\lambda}(I - Y_1(Y_1^HY_1 + Y_2^HY_2)^{-1}Y_1^H),\end{aligned}$$

and

$$\begin{aligned}\sin^2\Theta(X,Y) &= I - (Y_1^HY_1 + Y_2^HY_2)^{-\frac{1}{2}}Y_1^HY_1(Y_1^HY_1 + Y_2^HY_2)^{-\frac{1}{2}} \\ \Rightarrow \underline{\lambda}\big(\sin^2\Theta(X,Y)\big) &= \underline{\lambda}(I - Y_1^HY_1(Y_1^HY_1 + Y_2^HY_2)^{-1}) \\ &= \underline{\lambda}(I - Y_1(Y_1^HY_1 + Y_2^HY_2)^{-1}Y_1^H),\end{aligned}$$

where $\underline{\lambda}(\cdot)$ denotes the set of nonzero eigenvalues of a matrix. From (3.6) and (3.7), we deduce that if the nonzero singular values of $P_X^{\perp}(Y-X)(Y^HY)^{-\frac{1}{2}}$ are σ_1, σ_2, $\ldots$, then so are those of $\sin\Theta(X,Y)$, and moreover, those of $P_Y - P_X$ are σ_1, σ_1, σ_2, σ_2, $\ldots$, so

$$\|P_Y - P_X\|_2 = \|P_X^{\perp}(Y-X)(Y^HY)^{-\frac{1}{2}}\|_2 = \|\sin\Theta(X,Y)\|_2,$$
$$\frac{1}{\sqrt{2}}\|P_Y - P_X\|_F = \|P_X^{\perp}(Y-X)(Y^HY)^{-\frac{1}{2}}\|_F = \|\sin\Theta(X,Y)\|_F.$$

These prove the first equations in (3.5a) and (3.5b). By a symmetry argument, other equations in (3.5) also hold. □

COROLLARY 3.1. *The conditions and notations are as described in Theorem* 3.1. *Then there exists a permutation τ of* $\{1,2,\ldots,n\}$ *such that*

$$\text{(3.8)}\quad \max_{1\le j\le n}\rho\big((\alpha_j,\beta_j),(\widetilde{\alpha}_{\tau(j)},\widetilde{\beta}_{\tau(j)})\big) \le \min\left\{\|(\widetilde{Z}-Z)(\widetilde{Z}^H\widetilde{Z})^{-\frac{1}{2}}\|_2, \|(\widetilde{Z}-Z)(Z^HZ)^{-\frac{1}{2}}\|_2\right\}.$$

COROLLARY 3.2. *The conditions and notations are as described in Theorem* 3.2. *Then there exists a permutation μ of* $\{1,2,\ldots,n\}$ *such that*

$$\text{(3.9)}\quad \sqrt{\sum_{j=1}^{n}\left[\rho\big((\alpha_j,\beta_j),(\widetilde{\alpha}_{\mu(j)},\widetilde{\beta}_{\mu(j)})\big)\right]^2} \le \min\left\{\|(\widetilde{Z}-Z)(\widetilde{Z}^H\widetilde{Z})^{-\frac{1}{2}}\|_F, \|(\widetilde{Z}-Z)(Z^HZ)^{-\frac{1}{2}}\|_F\right\}.$$

The inequalities (3.1) and (3.3) correspond to two inequalities for perturbations of the ordinary SVD. Let A, $\widetilde{A} \in \mathbb{C}^{m\times n}$, with $\sigma(A) = \{\alpha_1,\ldots,\alpha_q\}$ and $\sigma(\widetilde{A}) = \{\widetilde{\alpha}_1,\ldots,\widetilde{\alpha}_q\}$, $q = \min\{m,n\}$. We also assume $\alpha_1 \ge \cdots \ge \alpha_q \ge 0$ and $\widetilde{\alpha}_1 \ge \cdots \ge \widetilde{\alpha}_q \ge 0$. Then

$$\text{(3.10)}\quad \max_{1\le j\le q}|\alpha_j - \widetilde{\alpha}_j| \le \|A-\widetilde{A}\|_2,$$

$$\text{(3.11)}\quad \sqrt{\sum_{j=1}^{q}|\alpha_j - \widetilde{\alpha}_j|^2} \le \|A-\widetilde{A}\|_F,$$

which are nothing but the inequalities (1.3) for the special unitarily invariant norms $\|\cdot\|_2$ and $\|\cdot\|_F$. The generalizations (3.1) and (3.3) of (3.10) and (3.11) are new. A few other generalizations have appeared in the literature. Sun [33] proved the following generalization of (3.10):

(3.12a)
$$\max_{1\le j\le n}\rho\big((\alpha_j,\beta_j),(\widetilde{\alpha}_{\tau(j)},\widetilde{\beta}_{\tau(j)})\big) \le \max_{\substack{x\in\mathbb{C}^n\\ \|x\|_2=1}}\rho\big((\|Ax\|_2,\|Bx\|_2),(\|\widetilde{A}x\|_2,\|\widetilde{B}x\|_2)\big)$$

$$\text{(3.12b)}\quad \le \|Z^{+}\|_2\|Z-\widetilde{Z}\|_2,$$

under the condition that

$$\max_{\substack{x\in\mathbb{C}^n\\ \|x\|_2=1}} \sqrt{\frac{\|(\widetilde{A}-A)x\|_2^2+\|(\widetilde{B}-B)x\|_2^2}{\|Ax\|_2^2+\|Bx\|_2^2}} < 1, \tag{3.13}$$

and for (3.11), Sun [30, Thm. 2.1] gave the following result:

$$\sqrt{1-\prod_{j=1}^{n}\left(1-\left[\rho\left(\left(\alpha_j,\beta_j\right),\left(\widetilde{\alpha}_{\tau(j)},\widetilde{\beta}_{\tau(j)}\right)\right)\right]^2\right)} \le \left\{1-\frac{|\det Z^H\widetilde{Z}|^2}{\det Z^HZ\,\det\widetilde{Z}^H\widetilde{Z}}\right\}^{\frac{1}{2}}, \tag{3.14a}$$

from which follows [30, Corollary 2.1]
(3.14b)

$$\sqrt{\sum_{j=1}^{n}\left[\rho\left(\left(\alpha_j,\beta_j\right),\left(\widetilde{\alpha}_{\tau(j)},\widetilde{\beta}_{\tau(j)}\right)\right)\right]^2} \le \left\{n\left[1-\sqrt[n]{\frac{|\det Z^H\widetilde{Z}|^2}{\det Z^HZ\,\det\widetilde{Z}^H\widetilde{Z}}}\right]\right\}^{\frac{1}{2}}.$$

Also, Paige [24], along the lines of [30], proved that if $\alpha_i^2+\beta_i^2=\widetilde{\alpha}_j^2+\widetilde{\beta}_j^2=1$, $i,j=1,2,\dots,n$, then there exists a permutation τ of $\{1,2,\dots,n\}$ such that

$$\sqrt{\sum_{j=1}^{n}\left[(\alpha_j-\widetilde{\alpha}_{\tau(j)})^2+(\beta_j-\widetilde{\beta}_{\tau(j)})^2\right]} \le \min_{Q\in\mathcal{U}_n}\|Z_0-\widetilde{Z}_0Q\|_F, \tag{3.15}$$

where $Z_0=Z(Z^HZ)^{-\frac{1}{2}}$ and $\widetilde{Z}_0=\widetilde{Z}(\widetilde{Z}^H\widetilde{Z})^{-\frac{1}{2}}$. As can be easily seen, the right-hand side of (3.15) involves a minimization over all $Q\in\mathcal{U}_n$. Thus it would be very difficult to compute, although the minimization may always provide a good estimation, so Paige [24] also gave (in our notations): if $\widetilde{Z}_1\in\mathbb{C}^{(m+p)\times(m+p-n)}$ such that $(\widetilde{Z}_0,\widetilde{Z}_1)\in\mathcal{U}_{m+p}$, then

$$\|\widetilde{Z}_1^HZ_0\|_F \le \min_{Q\in\mathcal{U}_n}\|Z_0-\widetilde{Z}_0Q\|_F \le \sqrt{2}\|\widetilde{Z}_1^HZ_0\|_F \le \sqrt{2}\|Z_0-\widetilde{Z}_0\|_F. \tag{3.16}$$

Now, we are in a position to compare those bounds with ours. First we consider the case for $\|\cdot\|_2$. When $m=p=n$, it is easy to see that (3.4) is the best one among all bounds. It is extremely difficult for us to relate the right-hand side of (3.12a) to that of (3.8) and to $\|\sin\Theta(Z,\widetilde{Z})\|_2$ as well. Therefore, we are unable here to say exactly which one of (3.12a) and (3.8) is better than the other and which one of (3.12a) and (3.1) is better as well. As to (3.12b), from

$$\|(Z^HZ)^{-\frac{1}{2}}\|_2 = \text{the smallest singular value of } Z = \|Z^+\|_2,$$
$$\|(\widetilde{Z}-Z)(Z^HZ)^{-\frac{1}{2}}\|_2 \le \|\widetilde{Z}-Z\|_2\|(Z^HZ)^{-\frac{1}{2}}\|_2 = \|Z^+\|_2\|\widetilde{Z}-Z\|_2,$$

it follows that (3.8) improves (3.12b), and so does (3.1).

For the case of $\|\cdot\|_F$, we could not compare (3.14a) to ours and to (3.15), but comparisons can be done if (3.14b) is used instead of (3.14a). A rather detailed comparison was conducted by Paige [24], whose conclusion is that the bound (3.15) (cf. (3.16)) is always just about as effective as the bound (3.14b). However, for large n the reverse cannot be said, i.e., for large n there is a possibility that the right-hand side of (3.14b) is quite large, while that of (3.15) remains small. Now we show that (3.3) and (3.15) are almost equivalent, but independent. To this end, we note [24, (2.11)]

$$\begin{aligned}(3.17)\qquad &\frac{1}{\sqrt{2}}\sqrt{\sum_{j=1}^{n}\left[(\alpha_j-\widetilde{\alpha}_{\tau(j)})^2+(\beta_j-\widetilde{\beta}_{\tau(j)})^2\right]}\\ &\le\sqrt{\sum_{j=1}^{n}\left[\rho\big((\alpha_j,\beta_j),(\widetilde{\alpha}_{\tau(j)},\widetilde{\beta}_{\tau(j)})\big)\right]^2}\\ &\le\sqrt{\sum_{j=1}^{n}\left[(\alpha_j-\widetilde{\alpha}_{\tau(j)})^2+(\beta_j-\widetilde{\beta}_{\tau(j)})^2\right]}.\end{aligned}$$

By Lemma 1.2, it follows that for any unitarily invariant norm $\|\cdot\|$

$$(3.18)\qquad \|\sin\Theta(Z,\widetilde{Z})\|=\|\widetilde{Z}_1^H Z_0\|.$$

Now, it will not be difficult for the reader, with (3.16)–(3.18) in mind, to derive a bound of the kind (3.15) via (3.3) and vice versa. The derived bound of one kind via the other is slightly weaker than its original by a factor $\sqrt{2}$. Comparing (3.9) with (3.15) seems to be somewhat troublesome. We end our comparisons with the following typical example [17], [31]: let $\{A,B\}$ be an (m,p,n)-GMP, and $\{\widetilde{A},\widetilde{B}\}=\{(1+r)A,(1+r)B\}$ its perturbed GMP. In this example, $\sigma\{A,B\}=\sigma\{\widetilde{A},\widetilde{B}\}$ and the right-hand sides of (3.1), (3.3), (3.4), (3.12a), (3.14), and (3.15) are all zero, thus those inequalities produce the best estimates. Those of (3.8), (3.9), and (3.12b) may be very large if r is chosen badly. This says that metrics on the Grassmann manifold have their own advantages in measuring the differences between two GMPs.

In short, we can say that our bounds are completely different from and at least as effective as, if not better than, those of Sun and Paige. Also, ours are formulated in an elegant way in very simple and intuitive terms.

Paige [24] and Sun [35] also included discussions on the case when $\{A,B\}$ is not an (m,p,n)-GMP, i.e., rank $\binom{A}{B}<n$. In my opinion, the GSVD problem for such a case should be regarded as an ill-conditioned one, as any small arbitrary perturbation to A and B may change the rank of $\binom{A}{B}$. The situation is quite similar to the generalized eigenvalue problem for singular matrix pencils [12], [13], [36]. Thus, without imposing any additional assumptions such as

$$\operatorname{rank}\begin{pmatrix}A\\B\end{pmatrix}=\operatorname{rank}\begin{pmatrix}\widetilde{A}\\\widetilde{B}\end{pmatrix},$$

it is almost impossible for us to develop a perfect perturbation theory for GSVD of $\{A,B\}$ with $\binom{A}{B}$ having no full column rank. The work done by Paige [24] and Sun [35] is indeed valuable for understanding the behavior of GSVD of a non-GMP in the presence of perturbations.

Remark 3.1. The reader may notice that we have used the same symbol τ in (3.1), (3.4), (3.8), (3.12), (3.14), and (3.15) for several permutations. This is not a result of our carelessness, but is intentional. Recall that ([24], [35]) all these τ's can be given explicitly, as we noted after (2.7). This is not the case for μ in (3.3) and (3.9). Generally, τ and μ are different (refer to the counterexample at the end of §2).

Remark 3.2. The bounds (3.10) and (3.11) can be obtained from (3.1) and (3.3) using a special limiting procedure. Such a procedure has also been used to derive the perturbation bounds for the standard eigenvalue problem and for the standard singular value problem via bounds for generalized problems; cf., e.g., Stewart [28], Sun [29], and Li [17]. We omit the derivations here.

We have given perturbation bounds in the two special unitarily invariant norms $\|\cdot\|_2$ and $\|\cdot\|_F$. Now, we consider the case when a general unitarily invariant norm is employed.

THEOREM 3.4. *The conditions and notations are as described in Theorem* 3.1. *Let*

$$(3.19)\qquad (\alpha_{n+i},\beta_{n+i}) = (-\alpha_i,\beta_i),\quad (\tilde{\alpha}_{n+j},\tilde{\beta}_{n+j}) = (-\tilde{\alpha}_j,\tilde{\beta}_j),\quad i,j=1,2,\dots,n.$$

Then there exists a permutation ν of $\{1,2,\dots,q\}$ such that for any unitarily invariant norm $\|\cdot\|$

$$(3.20)\qquad \left\|\mathrm{diag}\left(\rho\big((\alpha_1,\beta_1),(\tilde{\alpha}_{\nu(1)},\tilde{\beta}_{\nu(1)})\big),\dots,\rho\big((\alpha_q,\beta_q),(\tilde{\alpha}_{\nu(q)},\tilde{\beta}_{\nu(q)})\big)\right)\right\| \le \frac{\pi}{2}\|P_Z - P_{\tilde{Z}}\|,$$

where

$$q = 2n, \quad \textit{if } m \le n \textit{ and } p \le n;$$

$$\left\{\begin{array}{l} q = 2n + 2(p-n), \\ (\alpha_{2n+i},\beta_{2n+i}) = (\tilde{\alpha}_{2n+j},\tilde{\beta}_{2n+j}) = (1,0), \quad 1 \le i,j \le 2(p-n), \end{array}\right. \quad \textit{if } m \le n < p;$$

$$\left\{\begin{array}{l} q = 2n + 2(m-n), \\ (\alpha_{2n+i},\beta_{2n+i}) = (\tilde{\alpha}_{2n+j},\tilde{\beta}_{2n+j}) = (0,1), \quad 1 \le i,j \le 2(m-n), \end{array}\right. \quad \textit{if } m > n \ge p;$$

$$\left\{\begin{array}{ll} q = 2n + 2(p-n) + 2(m-n), & \\ (\alpha_{2n+i},\beta_{2n+i}) = (\tilde{\alpha}_{2n+j},\tilde{\beta}_{2n+j}) = (1,0), \quad 1 \le i,j \le 2(p-n), & \textit{if } m > n \\ (\alpha_{2n+2(p-n)+i},\beta_{2n+2(p-n)+i}) = (\tilde{\alpha}_{2n+2(p-n)+j},\tilde{\beta}_{2n+2(p-n)+j}) & \\ \qquad = (0,1), \quad 1 \le i,j \le 2(m-n), & \textit{and } p > n. \end{array}\right.$$

The appearance of (α_i,β_i) and $(\tilde{\alpha}_j,\tilde{\beta}_j)$ $(i,j > n)$ in this theorem makes it unsatisfactory, since the permutation ν may pair some of the (α_i,β_i) $(i \le n)$ to some of the $(\tilde{\alpha}_j,\tilde{\beta}_j)$ $(j > n)$. Therefore, it will be important to find a permutation ν with the property $\nu(i) \le n$, $i = 1,2,\dots,n$, such that (3.20) holds. On the other hand, when $\max\{m,p\} > n$ quite a few $(1,0)$ and/or $(0,1)$ that are not GSVs of the pairs considered intrude themselves into (3.20). This also makes (3.20) unsatisfactory. The reader will find in §4 that in the derivation of Theorems 3.1 and 3.3 there are also the same intruders $(1,0)$ and/or $(0,1)$. Fortunately, Propositions 2.2 and 2.3 help us to eliminate them. So, we intend to eliminate $(1,0)$ and/or $(0,1)$ in the right-hand side of (3.20) by choosing ν suitably. If we, indeed, could do all these, then Theorem 3.4 would become definitive. To this end, any answer (positive or negative) to the following conjecture would be helpful.

CONJECTURE. *Let* (α_i, β_i) *and* $(\widetilde{\alpha}_j, \widetilde{\beta}_j)$, $i, j = 1, 2, \ldots, n$, *be* $2n$ GSV, *and let* $(\alpha_{n+i}, \beta_{n+i})$ *and* $(\widetilde{\alpha}_{n+j}, \widetilde{\beta}_{n+j})$ $(i, j = 1, 2, \ldots, n)$ *be defined by* (3.19), *and* ν *be a permutation of* $\{1, 2, \ldots, 2n\}$. *Does there exist a permutation* μ *of* $\{1, 2, \ldots, n\}$ *such that for any unitarily invariant norm* $\|\cdot\|$,

$$\left\|\begin{pmatrix} \Sigma & \\ & \Sigma \end{pmatrix}\right\|$$
$$\leq \left\|\operatorname{diag}\left(\rho\left((\alpha_1, \beta_1), (\widetilde{\alpha}_{\nu(1)}, \widetilde{\beta}_{\nu(1)})\right), \ldots, \rho\left((\alpha_{2n}, \beta_{2n}), (\widetilde{\alpha}_{\nu(2n)}, \widetilde{\beta}_{\nu(2n)})\right)\right)\right\|,$$

where $\Sigma = \operatorname{diag}\left(\rho\left((\alpha_1, \beta_1), (\widetilde{\alpha}_{\mu(1)}, \widetilde{\beta}_{\mu(1)})\right), \ldots, \rho\left((\alpha_n, \beta_n), (\widetilde{\alpha}_{\mu(n)}, \widetilde{\beta}_{\mu(n)})\right)\right)$?

If there is a positive answer to this conjecture, elimination of $(1, 0)$ (or $(0, 1)$) from (3.20) becomes possible. This can be seen from (and by the proof of) Proposition 2.3.

We finish with a corollary of Theorem 3.4, but first we need the following lemma.

LEMMA 3.2. *The conditions and notations are as described in Lemma* 3.1. *Then for any unitarily invariant norm* $\|\cdot\|$,

$$\|P_Y - P_X\| \leq 2\left\|P_X^{\perp}(Y - X)(Y^H Y)^{-\frac{1}{2}}\right\| \tag{3.21a}$$
$$= 2\left\|P_Y^{\perp}(Y - X)(X^H X)^{-\frac{1}{2}}\right\|. \tag{3.21b}$$

Equation (3.21a) *becomes an identity if* $\|\cdot\|$ *is the trace class norm, i.e., the sum of all the singular values of a matrix.*

Proof. Let $\Phi(\cdot)$ be the symmetric gauge function associated with the $\|\cdot\|$. Then by Lemma 3.1, we have

$$\begin{aligned}
\|P_Y - P_X\| &= \Phi(\sigma_1, \sigma_1, \sigma_2, \sigma_2, \ldots) \\
&\leq \Phi(\sigma_1, 0, \sigma_2, 0, \ldots) + \Phi(0, \sigma_1, 0, \sigma_2, \ldots) \\
&= 2\left\|P_X^{\perp}(Y - X)(Y^H Y)^{-\frac{1}{2}}\right\| \\
&= 2\left\|P_Y^{\perp}(Y - X)(X^H X)^{-\frac{1}{2}}\right\|.
\end{aligned}$$

This establishes (3.21) and means that (3.21) is an equality if $\|\cdot\|$ is the trace class norm. □

COROLLARY 3.3. *The conditions and notations are as described in Theorem* 3.4. *Then*

$$\left\|\operatorname{diag}\left(\rho\left((\alpha_1, \beta_1), (\widetilde{\alpha}_{\nu(1)}, \widetilde{\beta}_{\nu(1)})\right), \ldots, \rho\left((\alpha_q, \beta_q), (\widetilde{\alpha}_{\nu(q)}, \widetilde{\beta}_{\nu(q)})\right)\right)\right\| \leq \begin{cases} \pi\left\|(\widetilde{Z} - Z)(\widetilde{Z}^H \widetilde{Z})^{-\frac{1}{2}}\right\|, \\ \pi\left\|(\widetilde{Z} - Z)(Z^H Z)^{-\frac{1}{2}}\right\|. \end{cases} \tag{3.22}$$

4. Proofs of Theorems 3.1–3.4. Before proving Theorems 3.1–3.3, we give five lemmas, the first four of which may also be of independent interest.

LEMMA 4.1. *Partition* $X \in \mathcal{U}_n$ *as*

$$X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}, \quad X_{11} \in \mathbb{C}^{k \times \ell}.$$

If $k+\ell>n$, then $\|X_{11}\|_2=1$.

Proof. It follows from $k+\ell>n$ that

$$\mathcal{R}\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} \bigcap \mathcal{R}\begin{pmatrix} I^{(k)} \\ 0 \end{pmatrix} \neq \emptyset,$$

in other words, there exist $g \in \mathbb{C}^\ell$, $h \in \mathbb{C}^k$ with $\|g\|_2=\|h\|_2=1$ such that

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} g = \begin{pmatrix} I^{(k)} \\ 0 \end{pmatrix} h \overset{\mathrm{def}}{=} x.$$

Obviously, $\|x\|_2=1$. On the other hand

$$X_{11} = (\, I^{(k)}, 0\,) \begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} \Rightarrow \|X_{11}\|_2 \geq h^H (\, I^{(k)}, 0\,) \begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} g = x^H x = 1.$$

This, together with the fact $\|X_{11}\|_2 \leq \|X\|_2 = 1$, leads to $\|X_{11}\|_2=1$. □

LEMMA 4.2. *Suppose that α_i, β_i, $\widetilde{\alpha}_j$, $\widetilde{\beta}_j \geq 0$, $\alpha_i^2+\beta_i^2=\widetilde{\alpha}_j^2+\widetilde{\beta}_j^2=1$, $i,j=1,2,\dots,n$, and set*

$$(4.1) \qquad \begin{cases} \Lambda = \mathrm{diag}\,(\alpha_1,\dots,\alpha_n), \\ \Omega = \mathrm{diag}\,(\beta_1,\dots,\beta_n), \end{cases} \quad \text{and} \quad \begin{cases} \widetilde{\Lambda} = \mathrm{diag}\,(\widetilde{\alpha}_1,\dots,\widetilde{\alpha}_n), \\ \widetilde{\Omega} = \mathrm{diag}\,(\widetilde{\beta}_1,\dots,\widetilde{\beta}_n). \end{cases}$$

Let U, V be two $n\times n$ unitary matrices. Then there exists a permutation τ of $\{1,2,\dots,n\}$ such that

$$(4.2) \qquad \max_{1\leq j\leq n} \rho\big(\,(\alpha_j,\beta_j),(\widetilde{\alpha}_{\tau(j)},\widetilde{\beta}_{\tau(j)})\,\big) \leq \|\Lambda U\widetilde{\Omega} - \Omega V\widetilde{\Lambda}\|_2.$$

Remark 4.1. An alternative formulation of (4.2) is

$$\min_{\mu}\max_{1\leq j\leq n} \rho\big(\,(\alpha_j,\beta_j),(\widetilde{\alpha}_{\mu(j)},\widetilde{\beta}_{\mu(j)})\,\big) = \min_{U,V\in\mathcal{U}_n} \|\Lambda U\widetilde{\Omega} - \Omega V\widetilde{\Lambda}\|_2.$$

This is because permutation matrices are also unitary.

A proof of Lemma 4.2 using Lemma 4.1 was implied in the proofs of Proposition 3.4 in Li [17]. But in the special case of Lemma 4.2, for convenience, we give an easy proof here.

Proof of Lemma 4.2. Without loss of generality, we assume that

$$(\alpha_1,\beta_1) \preceq \cdots \preceq (\alpha_n,\beta_n) \quad \text{and} \quad (\widetilde{\alpha}_1,\widetilde{\beta}_1) \preceq \cdots \preceq (\widetilde{\alpha}_n,\widetilde{\beta}_n).$$

Choose an index t satisfying

$$(4.3) \qquad \eta \overset{\mathrm{def}}{=} \max_{1\leq j\leq n} \rho\big(\,(\alpha_j,\beta_j),(\widetilde{\alpha}_j,\widetilde{\beta}_j)\,\big) = \rho\big(\,(\alpha_t,\beta_t),(\widetilde{\alpha}_t,\widetilde{\beta}_t)\,\big) > 0.$$

If $\eta=0$, then (4.2) becomes trivial. So we assume also $\eta>0$. The possible relative position of (α_t,β_t) and $(\widetilde{\alpha}_t,\widetilde{\beta}_t)$ on Γ is one of

$$(\alpha_t,\beta_t) \prec (\widetilde{\alpha}_t,\widetilde{\beta}_t), \qquad (\widetilde{\alpha}_t,\widetilde{\beta}_t) \prec (\alpha_t,\beta_t).$$

By a symmetry argument, it suffices to consider the first case, i.e., $(\alpha_t,\beta_t) \prec (\widetilde{\alpha}_t,\widetilde{\beta}_t)$. Denote by

$$\theta_i = \theta(\alpha_i,\beta_i), \quad \widetilde{\theta}_j = \theta(\widetilde{\alpha}_j,\widetilde{\beta}_j), \quad i,j = 1,2,\dots,n.$$

It is easy to verify that

$$\eta = \sin(\widetilde{\theta}_t - \theta_t) \tag{4.4}$$

and

$$\begin{cases} \min\limits_{1\le i\le t} \alpha_i = \alpha_t = \cos\theta_t, \\ \max\limits_{1\le i\le t} \beta_i = \beta_t = \sin\theta_t, \end{cases} \qquad \begin{cases} \max\limits_{t\le j\le n} \widetilde{\alpha}_j = \widetilde{\alpha}_t = \cos\widetilde{\theta}_t, \\ \min\limits_{t\le j\le n} \widetilde{\beta}_j = \widetilde{\beta}_t = \sin\widetilde{\theta}_t. \end{cases} \tag{4.5}$$

Next, we partition Λ, Ω, $\widetilde{\Lambda}$, $\widetilde{\Omega}$, U, and V as

$$\begin{aligned} \Lambda &= \mathrm{diag}\,(\Lambda_1,\,\Lambda_2), \quad \Omega = \mathrm{diag}\,(\Omega_1,\,\Omega_2), \quad \Lambda_1,\,\Omega_1 \in \mathbb{C}^{t\times t}, \\ \widetilde{\Lambda} &= \mathrm{diag}\,(\widetilde{\Lambda}_1,\,\widetilde{\Lambda}_2), \quad \widetilde{\Omega} = \mathrm{diag}\,(\widetilde{\Omega}_1,\,\widetilde{\Omega}_2), \quad \widetilde{\Lambda}_1,\,\widetilde{\Omega}_1 \in \mathbb{C}^{(t-1)\times(t-1)}, \end{aligned}$$

and

$$U = \begin{array}{c c} & \begin{array}{cc} t-1 & n-t+1 \end{array} \\ \begin{array}{c} t \\ n-t \end{array} & \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \end{array}, \qquad V = \begin{array}{c c} & \begin{array}{cc} t-1 & n-t+1 \end{array} \\ \begin{array}{c} t \\ n-t \end{array} & \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \end{array}.$$

Therefore,

$$E \overset{\mathrm{def}}{=} \Lambda U\widetilde{\Omega} - \Omega V\widetilde{\Lambda} = \begin{pmatrix} \Lambda_1 U_{11}\widetilde{\Omega}_1 - \Omega_1 V_{11}\widetilde{\Lambda}_1 & \Lambda_1 U_{12}\widetilde{\Omega}_2 - \Omega_1 V_{12}\widetilde{\Lambda}_2 \\ \Lambda_2 U_{21}\widetilde{\Omega}_1 - \Omega_2 V_{21}\widetilde{\Lambda}_1 & \Lambda_2 U_{22}\widetilde{\Omega}_2 - \Omega_2 V_{22}\widetilde{\Lambda}_2 \end{pmatrix},$$

which yields

$$\|E\|_2 \ge \|\Lambda_1 U_{12}\widetilde{\Omega}_2 - \Omega_1 V_{12}\widetilde{\Lambda}_2\|_2. \tag{4.6}$$

We will now prove that the right-hand side of (4.6) is greater than or equal to η. From (4.5) it follows that

$$\begin{cases} \|\Lambda_1^{-1}\|_2^{-1} = \min\limits_{1\le i\le t} \alpha_i = \cos\theta_t, \\ \|\Omega_1\|_2 = \max\limits_{1\le i\le t} \beta_i = \sin\theta_t, \end{cases} \qquad \begin{cases} \|\widetilde{\Lambda}_2\|_2 = \max\limits_{t\le j\le n} \widetilde{\alpha}_j = \cos\widetilde{\theta}_t, \\ \|\widetilde{\Omega}_2^{-1}\|_2^{-1} = \min\limits_{t\le j\le n} \widetilde{\beta}_j = \sin\widetilde{\theta}_t, \end{cases} \tag{4.7}$$

and by Lemma 4.1, it follows that

$$\begin{cases} U_{12},\, V_{12} \in \mathbb{C}^{t\times(n-t+1)} \\ t + (n-t+1) = n+1 > n \end{cases} \Rightarrow \|U_{12}\|_2 = \|V_{12}\|_2 = 1.$$

Combining (4.6) with (4.7) and the above equations produces

$$\begin{aligned} \|E\|_2 &\ge \|\Lambda_1 U_{12}\widetilde{\Omega}_2\|_2 - \|\Omega_1 V_{12}\widetilde{\Lambda}_2\|_2 \\ &\ge \|\Lambda_1^{-1}\|_2^{-1}\|U_{12}\|_2\|\widetilde{\Omega}_2^{-1}\|_2^{-1} - \|\Omega_1\|_2\|V_{12}\|_2\|\widetilde{\Lambda}_2\|_2 \\ &= \sin\widetilde{\theta}_t\cos\theta_t - \cos\widetilde{\theta}_t\sin\theta_t \\ &= \sin(\widetilde{\theta}_t - \theta_t). \end{aligned}$$

By (4.4) and (4.3), this is exactly what we need. □

LEMMA 4.3. *The conditions and notations are as described in Lemma* 4.2. *Then there exists a permutation μ of $\{1,2,\dots,n\}$ such that*

$$(4.8)\quad \sqrt{\sum_{j=1}^{n}\left[\rho\big((\alpha_j,\beta_j),(\widetilde{\alpha}_{\mu(j)},\widetilde{\beta}_{\mu(j)})\big)\right]^2} \le \frac{1}{\sqrt{2}}\left[\|\Lambda U\widetilde{\Omega}-\Omega V\widetilde{\Lambda}\|_F^2+\|\Lambda V\widetilde{\Omega}-\Omega U\widetilde{\Lambda}\|_F^2\right]^{\frac{1}{2}}.$$

Proof. Denote $U=(u_{ij})$, $V=(v_{ij})$. Then ($\Re\lambda$ denotes the real part of the complex number λ)

$$\begin{aligned}
&\|\Lambda U\widetilde{\Omega}-\Omega V\widetilde{\Lambda}\|_F^2+\|\Lambda V\widetilde{\Omega}-\Omega U\widetilde{\Lambda}\|_F^2\\
&=\sum_{i,j}\left[|\alpha_i\widetilde{\beta}_j u_{ij}-\beta_i\widetilde{\alpha}_j v_{ij}|^2+|\alpha_i\widetilde{\beta}_j v_{ij}-\beta_i\widetilde{\alpha}_j u_{ij}|^2\right]\\
&=\sum_{i,j}\left[(\alpha_i^2\widetilde{\beta}_j^2+\beta_i^2\widetilde{\alpha}_j^2)(|u_{ij}|^2+|v_{ij}|^2)-4\Re\alpha_i\beta_i\widetilde{\alpha}_j\widetilde{\beta}_j u_{ij}\bar{v}_{ij}\right]\\
&\ge\sum_{i,j}\left[(\alpha_i^2\widetilde{\beta}_j^2+\beta_i^2\widetilde{\alpha}_j^2)(|u_{ij}|^2+|v_{ij}|^2)-2\alpha_i\beta_i\widetilde{\alpha}_j\widetilde{\beta}_j(|u_{ij}|^2+|v_{ij}|^2)\right]\\
&=\sum_{i,j}(\alpha_i\widetilde{\beta}_j-\beta_i\widetilde{\alpha}_j)^2(|u_{ij}|^2+|v_{ij}|^2)\\
&\overset{\text{def}}{=}2\sum_{i,j}(\alpha_i\widetilde{\beta}_j-\beta_i\widetilde{\alpha}_j)^2 h_{ij},
\end{aligned}$$

where

$$H=(h_{ij})\overset{\text{def}}{=}\left(\frac{|u_{ij}|^2+|v_{ij}|^2}{2}\right)$$

is, evidently, a doubly stochastic matrix [22]. Thanks to the celebrated Birkhoff theorem: every doubly stochastic matrix is a convex combination of permutation matrices (see, e.g., Marshall and Olkin [22], [34, pp. 85–86]), and by a technique used frequently (see Hoffman and Wielandt [11] and Sun [29]), we can prove that there is a permutation μ of $\{1,2,\dots,n\}$ such that

$$\begin{aligned}
\|\Lambda U\widetilde{\Omega}-\Omega V\widetilde{\Lambda}\|_F^2+\|\Lambda V\widetilde{\Omega}-\Omega U\widetilde{\Lambda}\|_F^2 &\ge 2\sum_{j=1}^{n}(\alpha_j\widetilde{\beta}_{\mu(j)}-\beta_j\widetilde{\alpha}_{\mu(j)})^2\\
&=2\sum_{j=1}^{n}\left[\rho\big((\alpha_j,\beta_j),(\widetilde{\alpha}_{\mu(j)},\widetilde{\beta}_{\mu(j)})\big)\right]^2,
\end{aligned}$$

which yields (4.8) in a straightforward way. □

Remark 4.2. It can be seen from the above proof that (4.8) can be reversed in the sense that there is a permutation ν of $\{1,2,\dots,n\}$ such that

$$(4.9)\quad \frac{1}{\sqrt{2}}\left[\|\Lambda U\widetilde{\Omega}-\Omega V\widetilde{\Lambda}\|_F^2+\|\Lambda V\widetilde{\Omega}-\Omega U\widetilde{\Lambda}\|_F^2\right]^{\frac{1}{2}} \le \sqrt{\sum_{j=1}^{n}\left[\rho\big((\alpha_j,\beta_j),(-\widetilde{\alpha}_{\nu(j)},\widetilde{\beta}_{\nu(j)})\big)\right]^2}.$$

Readers may also find that the assumption $\alpha_i^2+\beta_i^2=\widetilde{\alpha}_j^2+\widetilde{\beta}_j^2=1$ is not essential. If this assumption, indeed, does not hold, it is sufficient to replace the left-hand side of (4.8) by

$$\sqrt{\sum_{j=1}^{n}|\alpha_j\widetilde{\beta}_{\mu(j)}-\beta_j\widetilde{\alpha}_{\mu(j)}|^2}$$

and the right-hand side of (4.9) by

$$\sqrt{\sum_{j=1}^{n}|\alpha_j\widetilde{\beta}_{\mu(j)}+\beta_j\widetilde{\alpha}_{\mu(j)}|^2}.$$

Then (4.8) and (4.9) with the modified left-hand side and right-hand side, respectively, remain valid.

Remark 4.3. Similarly to Remark 4.1, we see that Lemma 4.3 and Remark 4.2 yield the following results:

$$\min_{\mu}\sqrt{\sum_{j=1}^{n}\left[\rho\big(\,(\alpha_j,\beta_j),(\widetilde{\alpha}_{\mu(j)},\widetilde{\beta}_{\mu(j)})\,\big)\right]^2}$$
$$=\min_{U,V\in\mathcal{U}_n}\frac{1}{\sqrt{2}}\left[\|\Lambda U\widetilde{\Omega}-\Omega V\widetilde{\Lambda}\|_F^2+\|\Lambda V\widetilde{\Omega}-\Omega U\widetilde{\Lambda}\|_F^2\right]^{\frac{1}{2}}$$

and

$$\max_{U,V\in\mathcal{U}_n}\frac{1}{\sqrt{2}}\left[\|\Lambda U\widetilde{\Omega}-\Omega V\widetilde{\Lambda}\|_F^2+\|\Lambda V\widetilde{\Omega}-\Omega U\widetilde{\Lambda}\|_F^2\right]^{\frac{1}{2}}$$
$$=\max_{\nu}\sqrt{\sum_{j=1}^{n}\left[\rho\big(\,(\alpha_j,\beta_j),(-\widetilde{\alpha}_{\nu(j)},\widetilde{\beta}_{\nu(j)})\,\big)\right]^2}.$$

LEMMA 4.4. *The conditions and notations are as described in Lemma 4.2. Let $(\alpha_{n+i},\beta_{n+i})$ and $(\widetilde{\alpha}_{n+j},\widetilde{\beta}_{n+j})$ $(i,j=1,2,\dots,n)$ be defined by (3.19). Then there exists a permutation ν of $\{1,2,\dots,2n\}$ such that for any unitarily invariant norm $\|\cdot\|$,*

$$\big\|\mathrm{diag}\left(\rho\big(\,(\alpha_1,\beta_1),(\widetilde{\alpha}_{\nu(1)},\widetilde{\beta}_{\nu(1)})\,\big),\dots,\rho\big(\,(\alpha_{2n},\beta_{2n}),(\widetilde{\alpha}_{\nu(2n)},\widetilde{\beta}_{\nu(2n)})\,\big)\right)\big\|$$
$$\le\frac{\pi}{2}\big\|\mathrm{diag}\,(\Lambda U\widetilde{\Omega}-\Omega V\widetilde{\Lambda},\ \Lambda V\widetilde{\Omega}-\Omega U\widetilde{\Lambda})\big\|.$$

Proof. Set $I\equiv I^{(n)}$ and

$$\widehat{U}\overset{\mathrm{def}}{=}\frac{1}{\sqrt{2}}\begin{pmatrix}V&V\\U&-U\end{pmatrix},\quad\widehat{I}\overset{\mathrm{def}}{=}\frac{1}{\sqrt{2}}\begin{pmatrix}I&I\\I&-I\end{pmatrix},\quad\widetilde{I}\overset{\mathrm{def}}{=}\frac{1}{\sqrt{2}}\begin{pmatrix}I&-I\\I&I\end{pmatrix}.\tag{4.10}$$

An easy verification shows that $\widehat{U}$, $\widehat{I}$, $\widetilde{I}\in\mathcal{U}_{2n}$, $\widehat{I}^H=\widehat{I}^H$, $\widehat{I}^2=I^{(2n)}$, and

$$\widehat{I}\begin{pmatrix}&\Lambda\\\Lambda&\end{pmatrix}\widehat{I}=\begin{pmatrix}\Lambda&\\&-\Lambda\end{pmatrix},\quad\widehat{I}\begin{pmatrix}\Omega&\\&\Omega\end{pmatrix}\widehat{I}=\begin{pmatrix}\Omega&\\&\Omega\end{pmatrix}.\tag{4.11}$$

Note that

$$\begin{pmatrix} & \Lambda \\ \Lambda & \end{pmatrix} \widehat{U} \begin{pmatrix} \widetilde{\Omega} & \\ & \widetilde{\Omega} \end{pmatrix} - \begin{pmatrix} \Omega & \\ & \Omega \end{pmatrix} \widehat{U} \begin{pmatrix} \widetilde{\Lambda} & \\ & -\widetilde{\Lambda} \end{pmatrix}$$
$$= \operatorname{diag}(\Lambda U \widetilde{\Omega} - \Omega V \widetilde{\Lambda}, \Lambda V \widetilde{\Omega} - \Omega U \widetilde{\Lambda}) \widetilde{I},$$

so by (4.11),

$$\begin{aligned} &\begin{pmatrix} \Lambda & \\ & -\Lambda \end{pmatrix} \widehat{I} \widehat{U} \begin{pmatrix} \widetilde{\Omega} & \\ & \widetilde{\Omega} \end{pmatrix} - \begin{pmatrix} \Omega & \\ & \Omega \end{pmatrix} \widehat{I} \widehat{U} \begin{pmatrix} \widetilde{\Lambda} & \\ & -\widetilde{\Lambda} \end{pmatrix} \\ &= \widehat{I} \begin{pmatrix} & \Lambda \\ \Lambda & \end{pmatrix} \widehat{I}^2 \widehat{U} \begin{pmatrix} \widetilde{\Omega} & \\ & \widetilde{\Omega} \end{pmatrix} - \widehat{I} \begin{pmatrix} \Omega & \\ & \Omega \end{pmatrix} \widehat{I}^2 \widehat{U} \begin{pmatrix} \widetilde{\Lambda} & \\ & -\widetilde{\Lambda} \end{pmatrix} \\ &= \widehat{I} \operatorname{diag}(\Lambda U \widetilde{\Omega} - \Omega V \widetilde{\Lambda}, \Lambda V \widetilde{\Omega} - \Omega U \widetilde{\Lambda}) \widetilde{I}. \end{aligned}$$

Now, a result from Li [19] completes the proof. □

LEMMA 4.5 ([4, Chap. II, Thm. 5.1]). *Suppose $X \in \mathbb{C}^{n\times n}$ is partitioned as*

$$X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix},$$

then for any unitarily invariant norm $\|\cdot\|$ we have

$$\|X\| \geq \|\operatorname{diag}(X_{11}, X_{22})\|.$$

Now, we are ready for the proofs.

Suppose that (m,p,n)-GMP $\{A,B\}$ and $\{\widetilde{A},\widetilde{B}\}$ have the GSVDs

$$\left\{ \begin{array}{l} U^H A Q = \Sigma_A, \\ V^H B Q = \Sigma_B, \end{array} \right. \quad \text{and} \quad \left\{ \begin{array}{l} \widetilde{U}^H \widetilde{A} \widetilde{Q} = \Sigma_{\widetilde{A}}, \\ \widetilde{V}^H \widetilde{B} \widetilde{Q} = \Sigma_{\widetilde{B}}, \end{array} \right. \tag{4.12}$$

where Σ_A and Σ_B are defined by (1.5b) and (1.5c), $\Sigma_{\widetilde{A}}$ and $\Sigma_{\widetilde{B}}$ are also defined by (1.5b) and (1.5c), but with α_i, β_i replaced by $\widetilde{\alpha}_i$, $\widetilde{\beta}_i$, U, $\widetilde{U} \in \mathcal{U}_m$, and V, $\widetilde{V} \in \mathcal{U}_p$, Q, $\widetilde{Q} \in \mathbb{C}^{n\times n}$ are nonsingular. Suppose also, without loss of generality, that $\alpha_i^2 + \beta_i^2 = \widetilde{\alpha}_j^2 + \widetilde{\beta}_j^2 = 1$, $i,j = 1,2,\ldots,n$. Define Λ, Ω, $\widetilde{\Lambda}$, $\widetilde{\Omega}$ by (4.1).

We first show our theorems for the square case, i.e., $m = p = n$.

For this case $\Sigma_A = \Lambda$, $\Sigma_B = \Omega$, $\Sigma_{\widetilde{A}} = \widetilde{\Lambda}$, $\Sigma_{\widetilde{B}} = \widetilde{\Omega}$. It is easy to verify that

$$\begin{aligned} &B^H V U^H A - A^H U V^H B = Q^{-H} \Omega \Lambda Q^{-1} - Q^{-H} \Lambda \Omega Q^{-1} = 0 \\ &\Rightarrow (B^H V U^H A - A^H U V^H B) Z^+ = 0, \end{aligned}$$

therefore,

$$\begin{aligned} &-(B^H V U^H \widetilde{A} - A^H U V^H \widetilde{B}) \widetilde{Z}^+ \\ &= -(B^H V U^H \widetilde{A} - A^H U V^H \widetilde{B}) \widetilde{Z}^+ + (B^H V U^H A - A^H U V^H B) Z^+ \\ &= (B^H, -A^H) \begin{pmatrix} VU^H & \\ & UV^H \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} Z^+ - (B^H, -A^H) \begin{pmatrix} VU^H & \\ & UV^H \end{pmatrix} \begin{pmatrix} \widetilde{A} \\ \widetilde{B} \end{pmatrix} \widetilde{Z}^+ \\ &= (B^H, -A^H) \begin{pmatrix} VU^H & \\ & UV^H \end{pmatrix} (P_Z - P_{\widetilde{Z}}). \end{aligned}$$

Because $\operatorname{rank}\widetilde{Z} = n$, $\widetilde{Z}^{+}\widetilde{Z} = I^{(n)}$, postmultiplying both sides of the above equation by $\widetilde{Z}$ leads to

$$(4.13)\qquad B^H V U^H \widetilde{A} - A^H U V^H \widetilde{B} = -(B^H, -A^H)\begin{pmatrix} VU^H & \\ & UV^H \end{pmatrix}(P_Z - P_{\widetilde{Z}})\widetilde{Z}.$$

Inserting (4.12) into (4.13) we obtain

(4.14)

$$\Lambda V^H \widetilde{V}\widetilde{\Omega} - \Omega U^H \widetilde{U}\widetilde{\Lambda} = (\Omega, -\Lambda)\begin{pmatrix} U^H & \\ & V^H \end{pmatrix}(P_Z - P_{\widetilde{Z}})\begin{pmatrix} \widetilde{U} & \\ & \widetilde{V} \end{pmatrix}\begin{pmatrix} \widetilde{\Lambda} \\ \widetilde{\Omega} \end{pmatrix} \overset{\mathrm{def}}{=} E.$$

Now, by $V^H\widetilde{V} \in \mathcal{U}_n$, $U^H\widetilde{U} \in \mathcal{U}_n$, and Lemma 4.2, it follows that there is a permutation τ of $\{1, 2, \ldots, n\}$ such that

$$\|P_Z - P_{\widetilde{Z}}\|_2 \geq \|E\|_2 = \|\Lambda V^H \widetilde{V}\widetilde{\Omega} - \Omega U^H \widetilde{U}\widetilde{\Lambda}\|_2 \geq \max_{1\leq j\leq n} \rho\big((\alpha_j, \beta_j), (\widetilde{\alpha}_{\tau(j)}, \widetilde{\beta}_{\tau(j)}) \big),$$

which is (3.1).

To prove (3.3), we have, by interchanging the positions of $\{A, B\}$ and $\{\widetilde{A}, \widetilde{B}\}$ in (4.14), that

$$\widetilde{\Lambda}\widetilde{V}^H V \Omega - \widetilde{\Omega}\widetilde{U}^H U \Lambda = (\widetilde{\Omega}, -\widetilde{\Lambda})\begin{pmatrix} \widetilde{U}^H & \\ & \widetilde{V}^H \end{pmatrix}(P_{\widetilde{Z}} - P_Z)\begin{pmatrix} U & \\ & V \end{pmatrix}\begin{pmatrix} \Lambda \\ \Omega \end{pmatrix},$$

which gives

$$(4.15)\qquad \begin{aligned} &\Lambda U^H \widetilde{U}\widetilde{\Omega} - \Omega V^H \widetilde{V}\widetilde{\Lambda} \\ &\quad = (\Lambda, \Omega)\begin{pmatrix} U^H & \\ & V^H \end{pmatrix}(P_Z - P_{\widetilde{Z}})\begin{pmatrix} \widetilde{U} & \\ & \widetilde{V} \end{pmatrix}\begin{pmatrix} \widetilde{\Omega} \\ -\widetilde{\Lambda} \end{pmatrix} \overset{\mathrm{def}}{=} F. \end{aligned}$$

In the above derivation, the fact that P_Z and $P_{\widetilde{Z}}$ are Hermitian is employed (see, e.g., [34, p. 106]). Now, by $V^H\widetilde{V} \in \mathcal{U}_n$, $U^H\widetilde{U} \in \mathcal{U}_n$, and Lemma 4.3, it follows that there is a permutation μ of $\{1, 2, \ldots, n\}$ such that

$$(4.16)\qquad \|E\|_F^2 + \|F\|_F^2 \geq 2\sum_{j=1}^{n} \left[\rho\big((\alpha_j, \beta_j), (\widetilde{\alpha}_{\mu(j)}, \widetilde{\beta}_{\mu(j)}) \big)\right]^2.$$

We also have

(4.17)

$$\begin{aligned} \|E\|_F^2 + \|F\|_F^2 &\leq \left\|\begin{pmatrix} \Omega & -\Lambda \\ \Lambda & \Omega \end{pmatrix}\begin{pmatrix} U^H & \\ & V^H \end{pmatrix}(P_Z - P_{\widetilde{Z}})\begin{pmatrix} \widetilde{U} & \\ & \widetilde{V} \end{pmatrix}\begin{pmatrix} \widetilde{\Lambda} & \widetilde{\Omega} \\ \widetilde{\Omega} & -\widetilde{\Lambda} \end{pmatrix}\right\|_F^2 \\ &= \|P_Z - P_{\widetilde{Z}}\|_F^2, \end{aligned}$$

since

$$\begin{pmatrix} \Omega & -\Lambda \\ \Lambda & \Omega \end{pmatrix}, \quad \begin{pmatrix} \widetilde{\Lambda} & \widetilde{\Omega} \\ \widetilde{\Omega} & -\widetilde{\Lambda} \end{pmatrix} \in \mathcal{U}_{2n}.$$

Inequality (3.3) is a consequence of (4.16) and (4.17).

In order to prove (3.4), we note that

$$\begin{aligned} Z^H Z = A^H A + B^H B = Q^{-H} Q^{-1}, &\qquad \widetilde{Z}^H \widetilde{Z} = \widetilde{Q}^{-H} \widetilde{Q}^{-1} \\ \Rightarrow Q_1 = Q^{-1}(Z^H Z)^{-\frac{1}{2}} \in \mathcal{U}_n, &\qquad \widetilde{Q}_1 = \widetilde{Q}^{-1}(\widetilde{Z}^H \widetilde{Z})^{-\frac{1}{2}} \in \mathcal{U}_n. \end{aligned}$$

Therefore, from Lemma 4.2 and Proposition 2.2 (see Remark 2.1), it follows that there is a permutation τ of $\{1, 2, \dots, n\}$ independent of U_1 and V_1 such that

$$\begin{aligned} \|(Z^H Z)^{-\frac{1}{2}}(A^H V_1 \widetilde{B} - B^H U_1 \widetilde{A})(\widetilde{Z}^H \widetilde{Z})^{-\frac{1}{2}}\|_2 &= \|Q_1^H(\Lambda U^H V_1 \widetilde{V}\widetilde{\Omega} - \Omega V^H U_1 \widetilde{U}\widetilde{\Lambda})\widetilde{Q}_1\|_2 \\ &= \|\Lambda U^H V_1 \widetilde{V}\widetilde{\Omega} - \Omega V^H U_1 \widetilde{U}\widetilde{\Lambda}\|_2 \\ &\geq \max_{1 \leq j \leq n} \rho\big((\alpha_j, \beta_j), (\widetilde{\alpha}_{\tau(j)}, \widetilde{\beta}_{\tau(j)}) \big). \end{aligned}$$

Since U_1, V_1 are arbitrary unitary matrices, and the equality above holds for some special U_1 and V_1, we have proved (3.4).

To prove Theorem 3.4, it suffices to use Lemmas 4.4 and 4.5 with the two matrices E and F defined by (4.14) and (4.15), respectively, and noting that E and F are the two $n \times n$ diagonal blocks of the following $2n \times 2n$ matrix:

$$\begin{pmatrix} E & * \\ * & F \end{pmatrix} = \begin{pmatrix} \Omega & -\Lambda \\ \Lambda & \Omega \end{pmatrix} \begin{pmatrix} U^H & \\ & V^H \end{pmatrix} (P_Z - P_{\widetilde{Z}}) \begin{pmatrix} \widetilde{U} & \\ & \widetilde{V} \end{pmatrix} \begin{pmatrix} \widetilde{\Lambda} & \widetilde{\Omega} \\ \widetilde{\Omega} & -\widetilde{\Lambda} \end{pmatrix}.$$

So far, we have completed the proof for Theorem 3.3 and proved Theorems 3.1, 3.2, and 3.4 for the case of $m = p = n$.

Remark 4.4. We describe how (4.14) and (4.15) relate to the projective matrix spaces [26]. We note that the matrix space considered by [26] is all $n \times 2n$ complex matrices with full row rank n. Let $\{A, B\}$ and $\{\widetilde{A}, \widetilde{B}\}$ be two (n, n, n)-GMPs having GSVD as described above. Then Z and $\widetilde{Z}$ defined by (3.2) are two $2n \times n$ complex matrices with full column rank n, and Z^H and $\widetilde{Z}^H$ are two $n \times 2n$ complex matrices with full row rank n. We now want to show that the *chordal distance* $\chi(Z^H, \widetilde{Z}^H)$ introduced by Schwarz and Zaks [26] as a generalization of the chordal distance of the scalar case (refer to (1.8)) is nothing but $\|\sin\Theta(Z, \widetilde{Z})\|_2$; more specifically, we show

$$\chi(Z^H, \widetilde{Z}^H) = \|\Lambda V^H \widetilde{V}\widetilde{\Omega} - \Omega U^H \widetilde{U}\widetilde{\Lambda}\|_2 = \|\Lambda U^H \widetilde{U}\widetilde{\Omega} - \Omega V^H \widetilde{V}\widetilde{\Lambda}\|_2 \tag{4.18a}$$

$$= \|\sin\Theta(Z, \widetilde{Z})\|_2 = \|P_Z - P_{\widetilde{Z}}\|_2. \tag{4.18b}$$

To this end, we see by easy verification that

$$\begin{pmatrix} U\Lambda & -U\Omega \\ V\Omega & V\Lambda \end{pmatrix}, \quad \begin{pmatrix} \Lambda U^H & \Omega V^H \\ -\Omega U^H & \Lambda V^H \end{pmatrix}, \quad \begin{pmatrix} \widetilde{U}\widetilde{\Lambda} & -\widetilde{U}\widetilde{\Omega} \\ \widetilde{V}\widetilde{\Omega} & \widetilde{V}\widetilde{\Lambda} \end{pmatrix}, \quad \begin{pmatrix} \widetilde{\Lambda}\widetilde{U}^H & \widetilde{\Omega}\widetilde{V}^H \\ -\widetilde{\Omega}\widetilde{U}^H & \widetilde{\Lambda}\widetilde{V}^H \end{pmatrix}$$

are four $2n \times 2n$ unitary matrices. Thus, for any unitarily invariant norm $\|\cdot\|$, by Lemma 1.2 we have

$$\begin{aligned} \big\|\sin\Theta(Z, \widetilde{Z})\big\| &= \big\|\sin\Theta(ZQ, \widetilde{Z}\widetilde{Q})\big\| \qquad \text{(refer to (4.12))} \\ &= \begin{cases} \big\|\Lambda V^H \widetilde{V}\widetilde{\Omega} - \Omega U^H \widetilde{U}\widetilde{\Lambda}\big\|, \\ \big\|\Lambda U^H \widetilde{U}\widetilde{\Omega} - \Omega V^H \widetilde{V}\widetilde{\Lambda}\big\|. \end{cases} \end{aligned} \tag{4.19}$$

For $\|\cdot\| = \|\cdot\|_2$, this proves the first equation in (4.18b), while the second equation follows from Lemma 3.1. The equation (4.18a) follows immediately from the definition of $\chi(Z^H, \widetilde{Z}^H)$ in [26]. By the way, we point out an interesting result following from (4.18). The reader can easily deduce that

$$\text{(4.20)}\quad \|\Lambda V^H \widetilde{V}\widetilde{\Omega} - \Omega U^H \widetilde{U}\widetilde{\Lambda}\|_2 \le \|P_Z - P_{\widetilde{Z}}\|_2, \qquad \|\Lambda U^H \widetilde{U}\widetilde{\Omega} - \Omega V^H \widetilde{V}\widetilde{\Lambda}\|_2 \le \|P_Z - P_{\widetilde{Z}}\|_2.$$

In fact, we have already employed these two inequalities to prove (3.1). Equations (4.18a) and (4.18b) say that both equalities in (4.20) are always attained for some Z and $\widetilde{Z}$.

The rest of our proof is to reduce the case when A or B is not square to the square case that has just been proved above by augmenting the considered matrices suitably with zero block and unit matrices with appropriate dimensions and then applying our theorems for the square case together with Proposition 2.3. For instance, we consider the case: $m > n$ and $p > n$. The augumentation is done as follows:

$$\text{(4.21)}\quad \begin{aligned} \widehat{A} &= \left(\begin{array}{cc|c} I^{(p-n)} & & \\ & A & 0_{m+p-n,m-n} \end{array}\right), & \widehat{B} &= \left(\begin{array}{c|cc} & B & \\ 0_{m+p-n,p-n} & & I^{(m-n)} \end{array}\right), \\ \widehat{\widetilde{A}} &= \left(\begin{array}{cc|c} I^{(p-n)} & & \\ & \widetilde{A} & 0_{m+p-n,m-n} \end{array}\right), & \widehat{\widetilde{B}} &= \left(\begin{array}{c|cc} & \widetilde{B} & \\ 0_{m+p-n,p-n} & & I^{(m-n)} \end{array}\right), \end{aligned}$$

from which we have

$$\sigma\{\widehat{A}, \widehat{B}\} = \{\underbrace{(1,0),\dots,(1,0)}_{p-n}\} \bigcup \sigma\{A, B\} \bigcup \{\underbrace{(0,1),\dots,(0,1)}_{m-n}\},$$

$$\sigma\{\widehat{\widetilde{A}}, \widehat{\widetilde{B}}\} = \{\underbrace{(1,0),\dots,(1,0)}_{p-n}\} \bigcup \sigma\{\widetilde{A}, \widetilde{B}\} \bigcup \{\underbrace{(0,1),\dots,(0,1)}_{m-n}\}.$$

We leave to the reader to fill in the details.

We have completed the proofs of Theorems 3.1–3.4. □

Part II. Perturbation bounds for generalized singular subspaces. This part is organized as follows. We outline the background of the problem in §5, based on material from Sun [30, §1]. Our main results are discussed in §§6 and 8 according to the different features of the results. We will clarify these features in the two sections. The proofs of those results in §6 are given in detail in §7, while the proofs of those in §8 are only outlined in their own section.

5. Preliminaries to generalized singular subspaces.

Let $\{A, B\}$ be an (m, p, n)-GMP, whose GSVD is given by (1.5). We take a natural number ℓ satisfying

$$\text{(5.1)}\qquad \max\{n-p, 0\} < \ell < \min\{m, n\}.$$

From (1.5) we obtain formally

$$\text{(5.2a)}\quad \begin{aligned} A \begin{pmatrix} \overset{\ell}{Q_1}, & \overset{n-\ell}{Q_2} \end{pmatrix} &= \begin{pmatrix} \overset{\ell}{U_1}, & \overset{m-\ell}{U_2} \end{pmatrix} \times \begin{array}{c} \\ \ell \\ m-\ell \end{array}\!\! \begin{array}{c} \begin{array}{cc} \ell & n-\ell \end{array} \\ \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix} \end{array}, \\ B(Q_1, Q_2) &= \begin{pmatrix} \overset{p+\ell-n}{V_1}, & \overset{n-\ell}{V_2} \end{pmatrix} \times \begin{array}{c} \\ p+\ell-n \\ n-\ell \end{array}\!\! \begin{array}{c} \begin{array}{cc} \ell & n-\ell \end{array} \\ \begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix} \end{array}, \end{aligned}$$

and

$$
\begin{aligned}
A^H(U_1, U_2) &= \begin{array}{cc} \ell & n-\ell \\ \Big(P_1, & P_2\Big) \end{array} \times \begin{array}{c} \\ \ell \\ n-\ell \end{array} \!\!\begin{array}{cc} \ell & m-\ell \\ \left(\begin{matrix} A_{11}^H & 0 \\ 0 & A_{22}^H \end{matrix}\right. & \left.\right) \end{array}, \\
B^H(V_1, V_2) &= (P_1, P_2) \times \begin{array}{c} \\ \ell \\ n-\ell \end{array} \!\!\begin{array}{cc} p+\ell-n & n-\ell \\ \left(\begin{matrix} B_{11}^H & 0 \\ 0 & B_{22}^H \end{matrix}\right. & \left.\right) \end{array}.
\end{aligned} \tag{5.2b}
$$

Here (U_1, U_2) and (V_1, V_2) are unitary matrices; $Q = (Q_1, Q_2)$ and

$$
P = Q^{-H} = \begin{array}{cc} \ell & n-\ell \\ \Big(P_1, & P_2\Big) \end{array} \tag{5.3}
$$

are nonsingular matrices. Motivated by (5.2) and (5.3), Sun [30] suggested the following definition.

DEFINITION 5.1. Suppose that $\{A, B\}$ is an (m,p,n)-GMP and (5.1) holds for a natural number ℓ. Let $\mathcal{X}_1 \in \mathbb{C}^m$, $\mathcal{X}_3$ and $\mathcal{X}_4 \in \mathbb{C}^n$ be subspaces of dimension ℓ; let $\mathcal{X}_2 \in \mathbb{C}^p$ be a subspace of dimension $p + \ell - n$. Then $\{\mathcal{X}_j, j = 1,2,3,4\}$ forms a set of *generalized singular subspaces* (GSSSs) for $\{A, B\}$ if

$$
A\mathcal{X}_3 \subset \mathcal{X}_1, \quad B\mathcal{X}_3 \subset \mathcal{X}_2, \quad A^H\mathcal{X}_1 \subset \mathcal{X}_4, \quad B^H\mathcal{X}_2 \subset \mathcal{X}_4. \tag{5.4}
$$

Clearly, according to (5.2) and (5.3), the subspaces $\mathcal{X}_1 = \mathcal{R}(U_1)$, $\mathcal{X}_2 = \mathcal{R}(V_1)$, $\mathcal{X}_3 = \mathcal{R}(Q_1)$, and $\mathcal{X}_4 = \mathcal{R}(P_1)$ form a set of GSSSs for $\{A, B\}$. Further, [30] also proved that if there is a set $\{\mathcal{X}_j, j = 1,2,3,4\}$ of GSSSs of $\{A, B\}$ as described in Definition 5.1, then there exist unitary matrices $U = (U_1, U_2)$ and $V = (V_1, V_2)$ and a nonsingular matrix $Q = (Q_1, Q_2)$ with $\mathcal{X}_1 = \mathcal{R}(U_1)$, $\mathcal{X}_2 = \mathcal{R}(V_1)$, $\mathcal{X}_3 = \mathcal{R}(Q_1)$, and $\mathcal{X}_4 = \mathcal{R}(P_1)$ such that formally we have decompositions (5.2) and (5.3), where $\{A_{11}, B_{11}\}$ and $\{A_{22}, B_{22}\}$ are $(\ell, p + \ell - n, \ell)$-GMP and $(m - \ell, n - \ell, n - \ell)$-GMP, respectively.

In the sequel, we assume the (m,p,n)-GMP $\{A, B\}$ has the decomposition (1.5), and that (m,p,n)-GMP $\{\widetilde{A}, \widetilde{B}\}$ has the same decomposition with notation marked with tildes, e.g., corresponding to U_1, we have $\widetilde{U}_1$.

Note that in this paper we do not require the blocks A_{11}, B_{11}, ... in (5.2) to be diagonal as we did with the corresponding blocks in (1.5). Moreover, we do not require P_1 and Q_1 to have orthonormal columns as Sun did [30], we only require the completed matrices P and Q to be nonsingular.

Now, we define

$$
\Theta_1 \overset{\text{def}}{=} \Theta(U_1, \widetilde{U}_1), \quad \Theta_2 \overset{\text{def}}{=} \Theta(V_1, \widetilde{V}_1), \quad \Theta_3 \overset{\text{def}}{=} \Theta(Q_1, \widetilde{Q}_1), \quad \Theta_4 \overset{\text{def}}{=} \Theta(P_1, \widetilde{P}_1). \tag{5.5}
$$

By Lemma 1.2, we have

$$
\begin{aligned}
\|\sin\Theta_1\| &= \|\widetilde{U}_2^H U_1\|, & \|\sin\Theta_2\| &= \|\widetilde{V}_2^H V_1\|, \\
\|\sin\Theta_3\| &= \|\widetilde{P}_{20}^H Q_{10}\|, & \|\sin\Theta_4\| &= \|\widetilde{Q}_{20}^H P_{10}\|.
\end{aligned} \tag{5.6}
$$

Here

$$P_{j0} = P_j(P_j^H P_j)^{-\frac{1}{2}}, \quad Q_{j0} = Q_j(Q_j^H Q_j)^{-\frac{1}{2}}, \qquad j = 1, 2, \tag{5.7}$$

and thus $P_{j0}^H P_{j0} = Q_{j0}^H Q_{j0} = I$, similarly for $\widetilde{P}_{j0}$ and $\widetilde{Q}_{j0}$.

In the theorems of §6, as well as those of §8, there is an essential assumption—an assumption on the separations of two sets of GSVs.

Let Γ_1 and Γ_2 be two connected arcs on Γ which separate each other as shown in Fig. 2, where Γ_j is the arc corresponding to the angle θ_j, $j = 1, 2$. It is easy to verify the following.

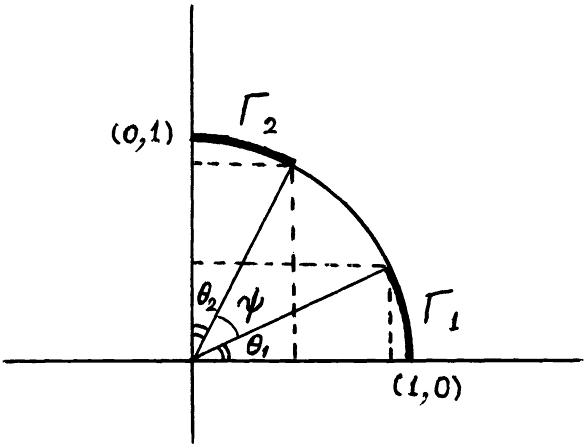


FIG. 2. *Separation.*

DESCRIPTION I (Li [17]). In Fig. 2,

$$\min_{\substack{(\alpha,\beta)\in\Gamma_1\\(\gamma,\delta)\in\Gamma_2}} \rho\big((\alpha,\beta),(\gamma,\delta)\big) = \sin\psi \overset{\mathrm{def}}{=} \eta. \tag{5.8}$$

Besides, there are two other descriptions.

'DESCRIPTION II. There exist $\alpha \geq 0$, $\delta > 0$, $\alpha + \delta \leq 1$, such that

$$\max_{w\in\Gamma_2} d(w,(0,1)\,) \leq \alpha \quad \text{and} \quad \min_{z\in\Gamma_1} d(z,(0,1)\,) \geq \alpha + \delta. \tag{5.9}$$

For Fig. 2, it suffices to choose $\alpha = \sin\theta_2$, $\alpha + \delta = \sin(\theta_2 + \psi)$. Description II was used by Sun [30] for establishing perturbation bounds for GSSSs.

DESCRIPTION III. There exist $\alpha \geq 0$, $\delta > 0$, $\alpha + \delta \leq 1$, such that

$$\max_{z\in\Gamma_1} d(z,(1,0)\,) \leq \alpha \quad \text{and} \quad \min_{w\in\Gamma_2} d(w,(1,0)\,) \geq \alpha + \delta. \tag{5.10}$$

For Fig. 2, it suffices to choose $\alpha = \sin\theta_1$, $\alpha + \delta = \sin(\theta_1 + \psi)$.

PROPOSITION 5.2. *The above three descriptions are equivalent. Moreover, if Description* II *or* III *holds, then Description* I *holds with*

$$\eta = (\alpha+\delta)\sqrt{1-\alpha^2} - \alpha\sqrt{1-(\alpha+\delta)^2}. \tag{5.11}$$

According to the standard sense for a distance between two sets with respect to a metric, η, as well as $\widetilde{\eta}$, $\widehat{\eta}$, and $\widetilde{\widetilde{\eta}}$, which will be introduced in §§6 and 8, is more reasonable than δ in describing gaps, and it would be better not to refer to α, δ in the final bounds. Although Sun's bounds [30] could be related to η without referring to α, δ (see our discussions at the end of §8), Sun [30] indeed used δ to describe it and regarded it as the gap between two sets of GSVs. It is easy to verify the following:

$$(2\alpha+\delta)\sqrt{1-(\alpha+\delta)^2} < (\alpha+\delta)\sqrt{1-\alpha^2} + \alpha\sqrt{1-(\alpha+\delta)^2} < (2\alpha+\delta)\sqrt{1-\alpha^2}$$
$$\Rightarrow \frac{1}{\sqrt{1-(\alpha+\delta)^2}} > \frac{2\alpha+\delta}{(\alpha+\delta)\sqrt{1-\alpha^2}+\alpha\sqrt{1-(\alpha+\delta)^2}} = \frac{\eta}{\delta} > \frac{1}{\sqrt{1-\alpha^2}}.$$

Given a set σ consisting of several GSVs, we hereafter prefer to use the notation $\sigma \subset \Gamma_1$ (or $\sigma \subset \Gamma_2$), which means that all points in Γ corresponding to elements of σ belong to Γ_1 (or to Γ_2).

6. Perturbation theorems for GSSSs.

In this section, we are only concerned with establishing bounds for unitarily invariant norms of $\sin\Theta_1$ and $\sin\Theta_2$.

THEOREM 6.1. *Let* $\{A,B\}$ *and* $\{\widetilde{A},\widetilde{B}\}$ *be two* (m,p,n)-GMP*s with the decompositions* (5.2) *and* (5.3), *and* $U = (U_1,U_2) \in \mathcal{U}_m$ *and* $V = (V_1,V_2) \in \mathcal{U}_p$. *Suppose that* $\sigma\{A_{11},B_{11}\} = \{(\alpha_i,\beta_i), i = 1,2,\dots,\ell\}$ *and* $\sigma\{\widetilde{A}_{22},\widetilde{B}_{22}\} = \{(\widetilde{\alpha}_j,\widetilde{\beta}_j), j = \ell+1,\dots,n\}$, *with* Γ_1 *and* Γ_2 *as shown in Fig.* 2 *and with* η *as defined by* (5.8). *If*

$$\sigma\{A_{11},B_{11}\} \subset \Gamma_1, \quad \sigma\{\widetilde{A}_{22},\widetilde{B}_{22}\} \subset \Gamma_2, \quad \eta > 0, \tag{6.1}$$

then for any unitarily invariant norm $\|\cdot\|$

$$\left\|\begin{pmatrix}\sin\Theta_1 & \\ & \sin\Theta_2\end{pmatrix}\right\| \le \frac{1}{\eta}\left\|\begin{pmatrix}U_1^H & \\ & V_1^H\end{pmatrix}(P_Z - P_{\widetilde{Z}})\begin{pmatrix}\widetilde{U}_2 & \\ & \widetilde{V}_2\end{pmatrix}\right\|, \tag{6.2}$$

$$\max_{i=1,2}\|\sin\Theta_i\| \le \frac{1}{\eta}\max\left\{\left\|\begin{pmatrix}U_1^H & \\ & V_1^H\end{pmatrix}(Z-\widetilde{Z})(\widetilde{Z}^H\widetilde{Z})^{-\frac{1}{2}}\right\|,\right.$$
$$\left.\left\|\begin{pmatrix}\widetilde{U}_2^H & \\ & \widetilde{V}_2^H\end{pmatrix}(Z-\widetilde{Z})(Z^HZ)^{-\frac{1}{2}}\right\|\right\}. \tag{6.3}$$

Here Z *and* $\widetilde{Z}$ *are defined by* (3.2).

Among all unitarily invariant norms, $\|\cdot\|_F$ often attracts additional attention due to its own properties. Here this is also the case.

THEOREM 6.2. *Let* $\{A,B\}$ *and* $\{\widetilde{A},\widetilde{B}\}$ *be two* (m,p,n)-GMP*s with the decompositions* (5.2) *and* (5.3), *and* $U = (U_1,U_2) \in \mathcal{U}_m$ *and* $V = (V_1,V_2) \in \mathcal{U}_p$. *Suppose that* $\sigma\{A_{11},B_{11}\} = \{(\alpha_i,\beta_i), i = 1,2,\dots,\ell\}$ *and* $\sigma\{\widetilde{A}_{22},\widetilde{B}_{22}\} = \{(\widetilde{\alpha}_j,\widetilde{\beta}_j), j = \ell+1,\dots,n\}$. *Let*

$$\underline{\eta} \overset{\text{def}}{=} \min_{\substack{1\le i\le \ell\\ \ell+1\le j\le n}} \rho\big((\alpha_i,\beta_i),(\widetilde{\alpha}_j,\widetilde{\beta}_j)\big), \tag{6.4a}$$

$$\eta_1 \overset{\text{def}}{=} \min_{\ell+1\le j\le n} \rho\big((1,0),(\widetilde{\alpha}_j,\widetilde{\beta}_j)\big) = \min_{\ell+1\le j\le n} \widetilde{\beta}_j, \tag{6.4b}$$

$$\eta_2 \overset{\text{def}}{=} \min_{1\le i\le \ell} \rho\big((\alpha_i,\beta_i),(0,1)\big) = \min_{1\le i\le \ell} \alpha_i, \tag{6.4c}$$

and

$$(6.5)\qquad \widetilde{\eta} \overset{\mathrm{def}}{=} \begin{cases} \underline{\eta} & \textit{if } m \le n \quad \textit{and} \quad p \le n, \\ \min\{\underline{\eta}, \eta_1\} & \textit{if } m \le n < p, \\ \min\{\underline{\eta}, \eta_2\} & \textit{if } m > n \ge p, \\ \min\{\underline{\eta}, \eta_1, \eta_2\} & \textit{if } m > n \quad \textit{and} \quad p > n. \end{cases}$$

Then if $\widetilde{\eta} > 0$, we have for the Frobenius norm $\|\cdot\|_F$,

$$(6.6)\qquad [\|\sin\Theta_1\|_F^2 + \|\sin\Theta_2\|_F^2]^{\frac{1}{2}} \le \frac{1}{\widetilde{\eta}} \left\| \begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix} (P_Z - P_{\widetilde{Z}}) \begin{pmatrix} \widetilde{U}_2 & \\ & \widetilde{V}_2 \end{pmatrix} \right\|_F,$$

$$(6.7)\qquad \begin{aligned} [\|\sin\Theta_1\|_F^2 + \|\sin\Theta_2\|_F^2]^{\frac{1}{2}} \le \frac{1}{\widetilde{\eta}} \Bigg[& \left\| \begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix} (Z - \widetilde{Z})(\widetilde{Z}^H \widetilde{Z})^{-\frac{1}{2}} \right\|_F^2 \\ & + \left\| \begin{pmatrix} \widetilde{U}_2^H & \\ & \widetilde{V}_2^H \end{pmatrix} (Z - \widetilde{Z})(Z^H Z)^{-\frac{1}{2}} \right\|_F^2 \Bigg]^{\frac{1}{2}}. \end{aligned}$$

Equations (6.4) and (6.5) also provide a description of the separation between two sets $\sigma\{A_{11}, B_{11}\}$ and $\sigma\{\widetilde{A}_{22}, \widetilde{B}_{22}\}$ of GSVs. This description is much weaker than that employed in Theorem 6.1 (cf. Descriptions I, II, and III in §5). In fact, if (6.1) holds with end points of Γ_1 and Γ_2 being elements of $\sigma\{A_{11}, B_{11}\}$ or $\sigma\{\widetilde{A}_{22}, \widetilde{B}_{22}\}$, then (6.4a) and (6.5) both hold with $\underline{\eta} = \widetilde{\eta} = \eta$, but the reverse cannot be said, because generally (6.4) and (6.5) in no way guarantee that there are Γ_1 and Γ_2 as shown in Fig. 2 such that (6.1) holds. The definitions (6.4) and (6.5) have some straightforward interpretations. Equation (6.4a) describes exactly how $\sigma\{A_{11}, B_{11}\}$ and $\sigma\{\widetilde{A}_{22}, \widetilde{B}_{22}\}$ separate in the sense of the chordal distance. Equations (6.4b), (6.4c), and (6.5) say that in some cases we should regard $(1,0)$ or $(0,1)$ as GSVs of $\sigma\{A_{11}, B_{11}\}$ or of $\sigma\{\widetilde{A}_{22}, \widetilde{B}_{22}\}$. In one way, the reader might think inclusion of $(1,0)$ or $(0,1)$ unsatisfactory, but we think it is quite reasonable in view of (5.2) as well as the GSVD described in Theorem 1.1. In fact, $(0,1) \in \sigma\{\widetilde{A}_{22}, \widetilde{B}_{22}\}$ if $m < n$, while $(1,0) \in \sigma\{A_{11}, B_{11}\}$ if $n > p$, and in giving η of Theorem 6.1, we indeed regard $(1,0) \in \sigma\{A_{11}, B_{11}\}$ and $(0,1) \in \sigma\{\widetilde{A}_{22}, \widetilde{B}_{22}\}$ as $(1,0) \in \Gamma_1$ and $(0,1) \in \Gamma_2$.

It can be seen that in the case of $\|\cdot\| = \|\cdot\|_F$, (6.6) is formally just (6.2), but the former is valid under much weaker conditions. The following theorem gives bounds in a general unitarily invariant norm in terms of $\widetilde{\eta}$. Here a constant enters into our bounds.

THEOREM 6.3. *The conditions and notations are as described in Theorem 6.2. Then there exists a constant $c \ge 1$ such that for any unitarily invariant norm $\|\cdot\|$ we have*

$$(6.8)\qquad \left\| \begin{pmatrix} \sin\Theta_1 & \\ & \sin\Theta_2 \end{pmatrix} \right\| \le c \cdot \frac{1}{\widetilde{\eta}} \left\| \begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix} (P_Z - P_{\widetilde{Z}}) \begin{pmatrix} \widetilde{U}_2 & \\ & \widetilde{V}_2 \end{pmatrix} \right\|,$$

$$(6.9)\qquad \begin{aligned} \left\| \begin{pmatrix} \sin\Theta_1 & \\ & \sin\Theta_2 \end{pmatrix} \right\| \le c \cdot \frac{1}{\widetilde{\eta}} \Bigg\{ & \left\| \begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix} (Z - \widetilde{Z})(\widetilde{Z}^H \widetilde{Z})^{-\frac{1}{2}} \right\| \\ & + \left\| \begin{pmatrix} \widetilde{U}_2^H & \\ & \widetilde{V}_2^H \end{pmatrix} (Z - \widetilde{Z})(Z^H Z)^{-\frac{1}{2}} \right\| \Bigg\}. \end{aligned}$$

We do not know what the smallest constant c that satisfies (6.8) and (6.9) is, but thanks to Bhatia, Davis, and Koosis [3], we shall prove that it can always be chosen so that

$$c \le \frac{\pi}{2}\int_0^{\pi} \frac{\sin t}{t}\,dt < 2.91. \tag{6.10}$$

Theorem 6.2 tells us that for the Frobenius norm the best constant c is 1.

Remark 6.1. The assumption that there must be positive gaps, say η, $\widetilde{\eta} > 0$, between $\sigma\{A_{11}, B_{11}\}$ and $\sigma\{\widetilde{A}_{22}, \widetilde{B}_{22}\}$ is essential. It is not difficult for the reader to find an example with $\eta = 0$ or $\widetilde{\eta} = 0$ for which small perturbations in A and/or B lead to big angles Θ_1 and Θ_2 (refer to Davis and Kahan [4] and Stewart [27], [34] for more intuitive insight on this matter).

7. Proofs of Theorems 6.1–6.3. Without loss of generality, let (m,p,n)-GMPs $\{A, B\}$ and $\{\widetilde{A}, \widetilde{B}\}$ have decompositions (5.2) and (5.3) with

$$\begin{pmatrix} A_{11} & \\ & A_{22} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} B_{11} & \\ & B_{22} \end{pmatrix}$$

being the matrices Σ_A and Σ_B, respectively, described in Theorem 1.1 and partitioned suitably, and with

$$\begin{pmatrix} \widetilde{A}_{11} & \\ & \widetilde{A}_{22} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \widetilde{B}_{11} & \\ & \widetilde{B}_{22} \end{pmatrix},$$

denoted by $\Sigma_{\widetilde{A}}$ and $\Sigma_{\widetilde{B}}$, being the matrices obtained from Σ_A and Σ_B by replacing α_j, β_j by $\widetilde{\alpha}_j$, $\widetilde{\beta}_j$, respectively. Otherwise, for example, because $\{A_{jj}, B_{jj}\}$ is GMP, we can find its GSVD, say $M_j^H A_{jj} L_j = \Sigma_{aj}$ and $N_j^H B_{jj} L_j = \Sigma_{bj}$, where M_j and N_j are unitary matrices with appropriate dimensions, and L_j is a nonsingular matrix, also with appropriate dimension. Thus (5.2) becomes

$$A(Q_1L_1, Q_2L_2) = (U_1M_1, U_2M_2)\begin{pmatrix} \Sigma_{a1} & \\ & \Sigma_{a2} \end{pmatrix},$$
$$B(Q_1L_1, Q_2L_2) = (V_1N_1, V_2N_2)\begin{pmatrix} \Sigma_{b1} & \\ & \Sigma_{b2} \end{pmatrix},$$

as required.

Set

$$\Sigma_{aj} = A_{jj}, \quad \Sigma_{bj} = B_{jj}, \quad \widetilde{\Sigma}_{aj} = \widetilde{A}_{jj}, \quad \widetilde{\Sigma}_{bj} = \widetilde{B}_{jj}, \tag{7.1}$$

and set

$$\begin{cases} \Lambda \equiv \operatorname{diag}(\Lambda_1, \Lambda_2) \overset{\text{def}}{=} \operatorname{diag}(\alpha_1, \dots, \alpha_n), \\ \Omega \equiv \operatorname{diag}(\Omega_1, \Omega_2) \overset{\text{def}}{=} \operatorname{diag}(\beta_1, \dots, \beta_n), \end{cases} \quad \Lambda_1, \Omega_1 \in \mathbb{C}^{\ell\times\ell}. \tag{7.2}$$

Similarly, we define $\widetilde{\Lambda}$, $\widetilde{\Omega}$, $\widetilde{\Lambda}_j$, $\widetilde{\Omega}_j$, $j = 1, 2$. It is easy to see that

$$\Lambda_1 = \Sigma_{a1}, \quad \widetilde{\Lambda}_1 = \widetilde{\Sigma}_{a1}, \quad \Omega_2 = \Sigma_{b2}, \quad \widetilde{\Omega}_2 = \widetilde{\Sigma}_{b2}. \tag{7.3}$$

We first consider the square case, i.e., $m = p = n$.

Now, besides (7.3), we also have $\Lambda_2 = \Sigma_{a2}$, $\widetilde{\Lambda}_2 = \widetilde{\Sigma}_{a2}$, $\Omega_1 = \Sigma_{b1}$, $\widetilde{\Omega}_1 = \widetilde{\Sigma}_{b1}$. Identity (4.14) from §4 and

$$(7.4)\quad \Lambda V^H \widetilde{V}\widetilde{\Omega} - \Omega U^H \widetilde{U}\widetilde{\Lambda} = \begin{pmatrix} \Lambda_1 V_1^H \widetilde{V}_1^H \widetilde{\Omega}_1 - \Omega_1 U_1 \widetilde{U}_1 \widetilde{\Lambda}_1 & \Lambda_1 V_1^H \widetilde{V}_2^H \widetilde{\Omega}_2 - \Omega_1 U_1 \widetilde{U}_2 \widetilde{\Lambda}_2 \\ \Lambda_2 V_2^H \widetilde{V}_1^H \widetilde{\Omega}_1 - \Omega_2 U_2 \widetilde{U}_1 \widetilde{\Lambda}_1 & \Lambda_2 V_2^H \widetilde{V}_2^H \widetilde{\Omega}_2 - \Omega_2 U_2 \widetilde{U}_2 \widetilde{\Lambda}_2 \end{pmatrix}$$

yields

$$(7.5\text{a})\quad \Lambda_1 V_1^H \widetilde{V}_2^H \widetilde{\Omega}_2 - \Omega_1 U_1^H \widetilde{U}_2 \widetilde{\Lambda}_2 = (\Omega_1, -\Lambda_1) \begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix} (P_Z - P_{\widetilde{Z}}) \begin{pmatrix} \widetilde{U}_2 & \\ & \widetilde{V}_2 \end{pmatrix} \begin{pmatrix} \widetilde{\Lambda}_2 \\ \widetilde{\Omega}_2 \end{pmatrix} \overset{\text{def}}{=} E.$$

Similarly, by (4.15) we have (cf. (7.4))

$$(7.5\text{b})\quad \Lambda_1 U_1^H \widetilde{U}_2^H \widetilde{\Omega}_2 - \Omega_1 V_1^H \widetilde{V}_2 \widetilde{\Lambda}_2 = (\Lambda_1, \Omega_1) \begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix} (P_Z - P_{\widetilde{Z}}) \begin{pmatrix} \widetilde{U}_2 & \\ & \widetilde{V}_2 \end{pmatrix} \begin{pmatrix} \widetilde{\Omega}_2 \\ -\widetilde{\Lambda}_2 \end{pmatrix} \overset{\text{def}}{=} F.$$

Similarly to (4.13) one may also get

$$B^H V U^H \widetilde{A} - A^H U V^H \widetilde{B} = -(B^H, -A^H) \begin{pmatrix} VU^H & \\ & UV^H \end{pmatrix} (Z - \widetilde{Z})$$

and thus

$$(7.6\text{a})\qquad \Lambda V^H \widetilde{V}\widetilde{\Omega} - \Omega U^H \widetilde{U}\widetilde{\Lambda} = (\Omega, -\Lambda) \begin{pmatrix} U^H & \\ & V^H \end{pmatrix} (Z - \widetilde{Z})(\widetilde{Z}^H \widetilde{Z})^{-\frac{1}{2}} \widehat{\widetilde{Q}}^H,$$

where $\widehat{\widetilde{Q}} = \widetilde{Q}^{-1}(\widetilde{Z}^H \widetilde{Z})^{-\frac{1}{2}} \in \mathcal{U}_n$ (refer to the proof of Theorem 3.3 in §4). Therefore

$$(7.7\text{a})\quad \Lambda_1 V_1^H \widetilde{V}_2 \widetilde{\Omega}_2 - \Omega_1 U_1^H \widetilde{U}_2 \widetilde{\Lambda}_2 = (\Omega_1, -\Lambda_1) \begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix} (Z - \widetilde{Z})(\widetilde{Z}^H \widetilde{Z})^{-\frac{1}{2}} \widehat{\widetilde{Q}}^H \begin{pmatrix} 0_{\ell, n-\ell} \\ I^{(n-\ell)} \end{pmatrix} \overset{\text{def}}{=} \widehat{E}.$$

By treating $\{A, B\}$ and $\{\widetilde{A}, \widetilde{B}\}$ equally, (7.6a) produces

$$\widetilde{\Lambda}\widetilde{V}^H V \Omega - \widetilde{\Omega}\widetilde{U}^H U \Lambda = (\widetilde{\Omega}, -\widetilde{\Lambda}) \begin{pmatrix} \widetilde{U}^H & \\ & \widetilde{V}^H \end{pmatrix} (\widetilde{Z} - Z)(Z^H Z)^{-\frac{1}{2}} \widehat{Q}^H$$

and thus

$$(7.6\text{b})\qquad \Lambda U^H \widetilde{U}\widetilde{\Omega} - \Omega V^H \widetilde{V}\widetilde{\Lambda} = \widehat{Q}(Z^H Z)^{-\frac{1}{2}}(Z - \widetilde{Z})^H \begin{pmatrix} \widetilde{U} & \\ & \widetilde{V} \end{pmatrix} \begin{pmatrix} \widetilde{\Omega} \\ -\widetilde{\Lambda} \end{pmatrix},$$

where $\widehat{Q} = Q^{-1}(Z^H Z)^{-\frac{1}{2}} \in \mathcal{U}_n$. So

$$(7.7\text{b})\quad \Lambda_1 U_1^H \widetilde{U}_2 \widetilde{\Omega}_2 - \Omega_1 V_1^H \widetilde{V}_2 \widetilde{\Lambda}_2 = (I^{(\ell)}, 0)\widehat{Q}(Z^H Z)^{-\frac{1}{2}}(Z - \widetilde{Z})^H \begin{pmatrix} \widetilde{U}_2^H & \\ & \widetilde{V}_2^H \end{pmatrix} \begin{pmatrix} \widetilde{\Omega}_2 \\ -\widetilde{\Lambda}_2 \end{pmatrix} \overset{\text{def}}{=} \widehat{F}$$

holds.

We are now ready for the proofs. First, we prove Theorem 6.1.

It follows from the assumption $(\alpha_i, \beta_i) \in \Gamma_1$, $i = 1, 2, \ldots, \ell$ and $(\tilde{\alpha}_j, \tilde{\beta}_j) \in \Gamma_2$, $j = \ell + 1, \ldots, n$ that

$$(7.8) \quad \begin{cases} \|\Lambda_1^{-1}\|_2^{-1} = \min\limits_{1 \le i \le \ell} |\alpha_i| \ge \cos\theta_1, \\ \|\Omega_1\|_2 = \max\limits_{1 \le i \le \ell} |\beta_i| \le \sin\theta_1, \end{cases} \quad \begin{cases} \|\tilde{\Omega}_2^{-1}\|_2^{-1} = \min\limits_{\ell+1 \le j \le n} |\tilde{\beta}_j| \ge \sin(\theta_1 + \psi), \\ \|\tilde{\Lambda}_2\|_2 = \max\limits_{\ell+1 \le j \le n} |\tilde{\alpha}_j| \le \cos(\theta_1 + \psi). \end{cases}$$

We claim that

$$(7.9) \quad \begin{aligned} \max\left\{\|\Lambda_1 V_1^H \tilde{V}_2 \tilde{\Omega}_2 - \Omega_1 U_1^H \tilde{U}_2 \tilde{\Lambda}_2\|, \|\Lambda_1 U_1^H \tilde{U}_2 \tilde{\Omega}_2 - \Omega_1 V_1^H \tilde{V}_2 \tilde{\Lambda}_2\|\right\} \\ \ge \eta \max\left\{\|V_1^H \tilde{V}_2\|, \|U_1^H \tilde{U}_2\|\right\} = \eta \max_{i=1,2} \|\sin\Theta_i\|. \end{aligned}$$

If (7.9) holds, then (6.3) is trivial. We have to prove (7.9). In fact, if $\|V_1^H \tilde{V}_2\| \ge \|U_1^H \tilde{U}_2\|$, then with the help of (7.8), we have

$$(7.10a) \quad \begin{aligned} &\|\Lambda_1 V_1^H \tilde{V}_2 \tilde{\Omega}_2 - \Omega_1 U_1^H \tilde{U}_2 \tilde{\Lambda}_2\| \\ &\quad \ge \|\Lambda_1 V_1^H \tilde{V}_2 \tilde{\Omega}_2\| - \|\Omega_1 U_1^H \tilde{U}_2 \tilde{\Lambda}_2\| \\ &\quad \ge \|\Lambda_1^{-1}\|_2^{-1} \|V_1^H \tilde{V}_2\| \|\tilde{\Omega}_2^{-1}\|_2^{-1} - \|\Omega_1\|_2 \|U_1^H \tilde{U}_2\| \|\tilde{\Lambda}_2\|_2 \\ &\quad \ge \|V_1^H \tilde{V}_2\| \sin\psi \\ &\quad = \eta \|\sin\Theta_2\|. \end{aligned}$$

If, on the other hand, $\|V_1^H \tilde{V}_2\| \le \|U_1^H \tilde{U}_2\|$, then again with the help of (7.8), we obtain

$$(7.10b) \quad \|\Lambda_1 U_1^H \tilde{U}_2 \tilde{\Omega}_2 - \Omega_1 V_1^H \tilde{V}_2 \tilde{\Lambda}_2\| \ge \eta \|\sin\Theta_1\|.$$

Combining inequalities (7.10a) and (7.10b) leads to (7.9). As in the proof of Lemma 4.4, to prove (6.2), we see that

$$(7.11) \quad \begin{pmatrix} \Lambda_1 & \\ & -\Lambda_1 \end{pmatrix} \hat{I} X \begin{pmatrix} \tilde{\Omega}_2 & \\ & \tilde{\Omega}_2 \end{pmatrix} - \begin{pmatrix} \Omega_1 & \\ & \Omega_1 \end{pmatrix} \hat{I} X \begin{pmatrix} \tilde{\Lambda}_2 & \\ & -\tilde{\Lambda} \end{pmatrix} = \hat{I} \begin{pmatrix} E & \\ & F \end{pmatrix} \tilde{I},$$

where unitary matrices $\hat{I}$ and $\tilde{I}$ are defined by (4.10) with I having appropriate dimensions, and

$$X = \frac{1}{\sqrt{2}} \begin{pmatrix} U_1^H \tilde{U}_2 & U_1^H \tilde{U}_2 \\ V_1^H \tilde{V}_2 & -V_1^H \tilde{V}_2 \end{pmatrix} = \begin{pmatrix} U_1^H \tilde{U}_2 & \\ & V_1^H \tilde{V}_2 \end{pmatrix} \hat{I}.$$

It is evident that (7.8) still holds with Λ_1, Ω_1, $\tilde{\Omega}_2$, and $\tilde{\Lambda}_2$ replaced by

$$\begin{pmatrix} \Lambda_1 & \\ & -\Lambda_1 \end{pmatrix}, \quad \begin{pmatrix} \Omega_1 & \\ & \Omega_1 \end{pmatrix}, \quad \begin{pmatrix} \tilde{\Omega}_2 & \\ & \tilde{\Omega}_2 \end{pmatrix}, \quad \begin{pmatrix} \tilde{\Lambda}_2 & \\ & -\tilde{\Lambda}_2 \end{pmatrix},$$

respectively. Therefore, as in (7.10a), one obtains from (7.11) that

$$\left\| \begin{pmatrix} U_1^H \widetilde{U}_2 & \\ & V_1^H \widetilde{V}_2 \end{pmatrix} \right\| \le \frac{1}{\eta} \left\| \begin{pmatrix} E & \\ & F \end{pmatrix} \right\|. \tag{7.12}$$

All the SVs of $U_1^H \widetilde{U}_2$ and those of $V_1^H \widetilde{V}_2$ are the same as those of $\sin\Theta_1$ and those of $\sin\Theta_2$, respectively, so the SVs of

$$\begin{pmatrix} U_1^H \widetilde{U}_2 & \\ & V_1^H \widetilde{V}_2 \end{pmatrix}$$

are the same as those of

$$\begin{pmatrix} \sin\Theta_1 & \\ & \sin\Theta_2 \end{pmatrix};$$

thus

$$\left\| \begin{pmatrix} U_1^H \widetilde{U}_2 & \\ & V_1^H \widetilde{V}_2 \end{pmatrix} \right\| \equiv \left\| \begin{pmatrix} \sin\Theta_1 & \\ & \sin\Theta_2 \end{pmatrix} \right\|. \tag{7.13}$$

It follows from Lemma 4.5 and

$$\begin{pmatrix} E & * \\ * & F \end{pmatrix} = \begin{pmatrix} \Omega_1 & -\Lambda_1 \\ \Lambda_1 & \Omega_1 \end{pmatrix} \begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix} (P_Z - P_{\widetilde{Z}}) \begin{pmatrix} \widetilde{U}_2 & \\ & \widetilde{V}_2 \end{pmatrix} \begin{pmatrix} \widetilde{\Lambda}_2 & \widetilde{\Omega}_2 \\ \widetilde{\Omega}_2 & -\widetilde{\Lambda}_2 \end{pmatrix}$$

and

$$\begin{pmatrix} \Omega_1 & -\Lambda_1 \\ \Lambda_1 & \Omega_1 \end{pmatrix} \in \mathcal{U}_{2\ell}, \qquad \begin{pmatrix} \widetilde{\Lambda}_2 & \widetilde{\Omega}_2 \\ \widetilde{\Omega}_2 & -\widetilde{\Lambda}_2 \end{pmatrix} \in \mathcal{U}_{2(n-\ell)}$$

that

$$\left\| \begin{pmatrix} E & \\ & F \end{pmatrix} \right\| \le \left\| \begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix} (P_Z - P_{\widetilde{Z}}) \begin{pmatrix} \widetilde{U}_2 & \\ & \widetilde{V}_2 \end{pmatrix} \right\|. \tag{7.14}$$

Equation (6.2) is a consequence of (7.12), (7.13), and (7.14).

Now, we prove Theorem 6.2. Denote by $V_1^H \widetilde{V}_2 = (v_{ij})$, $U_1^H \widetilde{U}_2 = (u_{ij})$, then,

$$\begin{aligned} &\|\Lambda_1 V_1^H \widetilde{V}_2 \widetilde{\Omega}_2 - \Omega_1 U_1^H \widetilde{U}_2 \widetilde{\Lambda}_2\|_F^2 + \|\Lambda_1 U_1^H \widetilde{U}_2 \widetilde{\Omega}_2 - \Omega_1 V_1^H \widetilde{V}_2 \widetilde{\Lambda}_2\|_F^2 \\ &\quad = \sum_{i,j} \left[|\alpha_i \widetilde{\beta}_j v_{ij} - \beta_i \widetilde{\alpha}_j u_{ij}|^2 + |\alpha_i \widetilde{\beta}_j u_{ij} - \beta_i \widetilde{\alpha}_j v_{ij}|^2 \right] \\ &\quad = \sum_{i,j} \left[(\alpha_i^2 \widetilde{\beta}_j^2 + \beta_i^2 \widetilde{\alpha}_j^2)(|u_{ij}|^2 + |v_{ij}|^2) - 4\Re \alpha_i \beta_i \widetilde{\alpha}_j \widetilde{\beta}_j u_{ij} \bar{v}_{ij} \right] \\ &\quad \ge \sum_{i,j} (\alpha_i \widetilde{\beta}_j - \beta_i \widetilde{\alpha}_j)^2 (|u_{ij}|^2 + |v_{ij}|^2) \\ &\quad \ge \eta^2 (\|V_1^H \widetilde{V}_2\|_F^2 + \|U_1^H \widetilde{U}_2\|_F^2). \end{aligned} \tag{7.15}$$

Equation (6.7) follows in a straightforward way from (7.7) and (7.15). Identities (7.5a) and (7.5b) and inequalities (7.14) and (7.15) then yield (6.6).

We now prove Theorem 6.3 for the square case. Again, (7.11) is the key. For the fact that α_i, β_i, $\widetilde{\alpha}_j$, $\widetilde{\beta}_j \geq 0$, we easily see that

$$\rho\big((-\alpha_i,\beta_i),(\widetilde{\alpha}_j,\widetilde{\beta}_j)\big) = \rho\big((\alpha_i,\beta_i),(-\widetilde{\alpha}_j,\widetilde{\beta}_j)\big) \geq \rho\big((\alpha_i,\beta_i),(\widetilde{\alpha}_j,\widetilde{\beta}_j)\big) \geq \eta$$

for $1 \leq i \leq \ell$, $\ell+1 \leq j \leq n$. So from Li [18, §4, Lemma 4.1], which was proved by the author using results due to [1] and [3], it follows that there exists a constant $c \geq 1$, which can be chosen such that (6.10) holds, so that (note $\widehat{E} = E$ and $\widehat{F} = F$)

$$\left\|\begin{pmatrix} U_1^H\widetilde{U}_2 & \\ & V_1^H\widetilde{V}_2 \end{pmatrix}\right\| \leq c\cdot\frac{1}{\eta}\left\|\begin{pmatrix} E & \\ & F \end{pmatrix}\right\| = c\cdot\frac{1}{\eta}\left\|\begin{pmatrix} \widehat{E} & \\ & \widehat{F} \end{pmatrix}\right\|.$$

This together with (7.13) and (7.14) yields (6.8), and with (7.13),

$$\left\|\begin{pmatrix} \widehat{E} & \\ & \widehat{F} \end{pmatrix}\right\| \leq \left\|\begin{pmatrix} \widehat{E} & \\ & 0 \end{pmatrix}\right\| + \left\|\begin{pmatrix} 0 & \\ & \widehat{F} \end{pmatrix}\right\| = \big\|\widehat{E}\big\| + \big\|\widehat{F}\big\|$$

and (7.7) produce (6.9).

We treat the case when A or B is not square in the same way as we did in §4. For an example, suppose that $m > n$ and $p > n$. Let $\widehat{A}$, $\widehat{B}$, $\widehat{\widetilde{A}}$, $\widehat{\widetilde{B}}$ be as defined by (4.21). It is easy to verify that

$$\widehat{U}^H\widehat{A}\widehat{Q} \overset{\text{def}}{=} \begin{pmatrix} I & \\ & U^H \end{pmatrix}\widehat{A}\begin{pmatrix} I & & \\ & Q & \\ & & I \end{pmatrix} = \begin{pmatrix} I & & \\ & \Lambda_1 & \\ & & \Sigma_{a2},\ 0 \end{pmatrix} \overset{\text{def}}{=} \begin{pmatrix} \widehat{\Lambda}_1 & \\ & \widehat{\Lambda}_2 \end{pmatrix},$$

$$\widehat{V}^H\widehat{B}\widehat{Q} \overset{\text{def}}{=} \begin{pmatrix} V^H & \\ & I \end{pmatrix}\widehat{B}\widehat{Q} = \begin{pmatrix} 0,\ \Sigma_{b1} & & \\ & \Omega_2 & \\ & & I \end{pmatrix} \overset{\text{def}}{=} \begin{pmatrix} \widehat{\Omega}_1 & \\ & \widehat{\Omega}_2 \end{pmatrix},$$

where

$$\widehat{\Lambda}_1 = \begin{pmatrix} I & \\ & \Lambda_1 \end{pmatrix},\quad \widehat{\Lambda}_2 = (\Sigma_{a2},\ 0),\quad \widehat{\Omega}_1 = (0,\ \Sigma_{b1}),\quad \widehat{\Omega}_2 = \begin{pmatrix} \Omega_2 & \\ & I \end{pmatrix},$$

and

$$\widehat{\widetilde{U}}^H\widehat{\widetilde{A}}\widehat{\widetilde{Q}} \overset{\text{def}}{=} \begin{pmatrix} I & \\ & \widetilde{U}^H \end{pmatrix}\widehat{\widetilde{A}}\begin{pmatrix} I & & \\ & \widetilde{Q} & \\ & & I \end{pmatrix} = \begin{pmatrix} I & & \\ & \widetilde{\Lambda}_1 & \\ & & \widetilde{\Sigma}_{a2},\ 0 \end{pmatrix} \overset{\text{def}}{=} \begin{pmatrix} \widehat{\widetilde{\Lambda}}_1 & \\ & \widehat{\widetilde{\Lambda}}_2 \end{pmatrix},$$

$$\widehat{\widetilde{V}}^H\widehat{\widetilde{B}}\widehat{\widetilde{Q}} \overset{\text{def}}{=} \begin{pmatrix} \widetilde{V}^H & \\ & I \end{pmatrix}\widehat{\widetilde{B}}\widehat{\widetilde{Q}} = \begin{pmatrix} 0,\ \widetilde{\Sigma}_{b1} & & \\ & \widetilde{\Omega}_2 & \\ & & I \end{pmatrix} \overset{\text{def}}{=} \begin{pmatrix} \widehat{\widetilde{\Omega}}_1 & \\ & \widehat{\widetilde{\Omega}}_2 \end{pmatrix},$$

where

$$\widehat{\widetilde{\Lambda}}_1 = \begin{pmatrix} I & \\ & \widetilde{\Lambda}_1 \end{pmatrix},\quad \widehat{\widetilde{\Lambda}}_2 = (\widetilde{\Sigma}_{a2},\ 0),\quad \widehat{\widetilde{\Omega}}_1 = (0,\ \widetilde{\Sigma}_{b1}),\quad \widehat{\widetilde{\Omega}}_2 = \begin{pmatrix} \widetilde{\Omega}_2 & \\ & I \end{pmatrix}.$$

Moreover,

$$\sigma\{\widehat{\Lambda}_1,\widehat{\Omega}_1\} = \{\underbrace{(1,0),\dots,(1,0)}_{p-n}\}\bigcup\sigma\{A_{11},B_{11}\},$$

$$\sigma\{\widehat{\widetilde{\Lambda}}_2,\widehat{\widetilde{\Omega}}_2\} = \sigma\{\widetilde{A}_{22},\widetilde{B}_{22}\}\bigcup\{\underbrace{(0,1),\dots,(0,1)}_{m-n}\}.$$

Therefore, if the conditions of Theorem 6.1 or those of Theorem 6.2 hold for $\{A,B\}$ and $\{\widetilde{A},\widetilde{B}\}$, they also hold for $(m+p-n, m+p-n, m+p-n)$-GMP $\{\widehat{A},\widehat{B}\}$ and $\{\widehat{\widetilde{A}},\widehat{\widetilde{B}}\}$ with $\widehat{\ell} = p-n+\ell$. The reader must finish the details.

8. More perturbation theorems for GSSSs with comparisons. In this section, we present not only more bounds for $\sin\Theta_1$ and $\sin\Theta_2$, but also bounds for $\sin\Theta_3$ and $\sin\Theta_4$.

A very interesting feature of Theorems 6.1–6.3 is that none of the bounds for $\sin\Theta_1$ and $\sin\Theta_2$ bear any relation to P and Q, or to Θ_3 and Θ_4. This is not the case for the theorems we will develop below. Another different feature is that here we start by defining four residual matrices due to Sun [30] instead of (7.5)–(7.7) in §7. But here, our derivation of perturbation equations from the four residuals seems more elegant.

Assume now that we have (5.2) and (5.3) for $\{A, B\}$ and $\{\widetilde{A}, \widetilde{B}\}$. Define

$$\begin{aligned} R_A &\overset{\mathrm{def}}{=} \widetilde{A}Q_1 - U_1A_{11} = (\widetilde{A} - A)Q_1, \\ R_B &\overset{\mathrm{def}}{=} \widetilde{B}Q_1 - V_1B_{11} = (\widetilde{B} - B)Q_1, \\ R_{A^H} &\overset{\mathrm{def}}{=} \widetilde{A}^HU_1 - P_1A_{11}^H = (\widetilde{A} - A)^HU_1, \\ R_{B^H} &\overset{\mathrm{def}}{=} \widetilde{B}^HV_1 - P_1B_{11}^H = (\widetilde{B} - B)^HV_1. \end{aligned} \tag{8.1}$$

Suppose more strongly that (5.2) and (5.3) are as described in the beginning of §7 up to (7.3). Assume for the moment that $m = p = n$. Then

$$A_{jj} = \Lambda_j,\ B_{jj} = \Omega_j, \quad \widetilde{A}_{jj} = \widetilde{\Lambda}_j,\ \widetilde{B}_{jj} = \widetilde{\Omega}_j, \quad j = 1, 2. \tag{8.2}$$

From (8.1) it follows that

$$\begin{aligned} \underline{R}_A &\overset{\mathrm{def}}{=} \widetilde{U}_2^HR_A = \widetilde{\Lambda}_2\widetilde{P}_2^HQ_1 - \widetilde{U}_2^HU_1\Lambda_1, \\ \underline{R}_B &\overset{\mathrm{def}}{=} \widetilde{V}_2^HR_B = \widetilde{\Omega}_2\widetilde{P}_2^HQ_1 - \widetilde{V}_2^HV_1\Omega_1, \\ \underline{R}_{A^H} &\overset{\mathrm{def}}{=} \widetilde{Q}_2^HR_{A^H} = \widetilde{\Lambda}_2\widetilde{U}_2^HU_1 - \widetilde{Q}_2^HP_1\Lambda_1, \\ \underline{R}_{B^H} &\overset{\mathrm{def}}{=} \widetilde{Q}_2^HR_{B^H} = \widetilde{\Omega}_2\widetilde{V}_2^HV_1 - \widetilde{Q}_2^HP_1\Omega_1. \end{aligned} \tag{8.3}$$

Therefore, we get

$$C_1 \overset{\mathrm{def}}{=} -\widetilde{\Omega}_2\underline{R}_A + \widetilde{\Lambda}_2\underline{R}_B = \widetilde{\Omega}_2\widetilde{U}_2^HU_1\Lambda_1 - \widetilde{\Lambda}_2\widetilde{V}_2^HV_1\Omega_1, \tag{8.4a}$$

$$C_2 \overset{\mathrm{def}}{=} \underline{R}_{A^H}\Omega_1 - \underline{R}_{B^H}\Lambda_1 = \widetilde{\Lambda}_2\widetilde{U}_2^HU_1\Omega_1 - \widetilde{\Omega}_2\widetilde{V}_2^HV_1\Lambda_1, \tag{8.4b}$$

$$\begin{aligned} C_3 &\overset{\mathrm{def}}{=} \widetilde{\Lambda}_2\underline{R}_A + \underline{R}_{A^H}\Lambda_1 = \widetilde{\Lambda}_2^2\widetilde{P}_2^HQ_1 - \widetilde{Q}_2^HP_1\Lambda_1^2, \\ C_4 &\overset{\mathrm{def}}{=} \widetilde{\Omega}_2\underline{R}_B + \underline{R}_{B^H}\Omega_1 = \widetilde{\Omega}_2^2\widetilde{P}_2^HQ_1 - \widetilde{Q}_2^HP_1\Omega_1^2, \end{aligned}$$

$$C_5 \overset{\mathrm{def}}{=} C_3\Omega_1^2 - C_4\Lambda_1^2 = \widetilde{\Lambda}_2^2\widetilde{P}_2^HQ_1\Omega_1^2 - \widetilde{\Omega}_2^2\widetilde{P}_2^HQ_1\Lambda_1^2, \tag{8.5a}$$

$$C_6 \overset{\mathrm{def}}{=} -\widetilde{\Omega}_2^2C_3 + \widetilde{\Lambda}_2^2C_4 = \widetilde{\Omega}_2^2\widetilde{Q}_2^HP_1\Lambda_1^2 - \widetilde{\Lambda}_2^2\widetilde{Q}_2^HP_1\Omega_1^2. \tag{8.5b}$$

THEOREM 8.1. *The conditions and notations are as described in Theorem* 6.1. *Then*

$$\begin{aligned} \max_{i=1,2}\|\sin\Theta_i\| \le \frac{1}{\eta}\max\Bigg\{ &\left\|\begin{pmatrix} \widetilde{U}_2^H & \\ & \widetilde{V}_2^H \end{pmatrix}(Z - \widetilde{Z})Q_{10}\right\| \|Z^+\|_2, \\ &\left\|\begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix}(Z - \widetilde{Z})\widetilde{Q}_{20}\right\| \|\widetilde{Z}^+\|_2\Bigg\}, \end{aligned} \tag{8.6}$$

and

(8.7)

$$\left\|\sin\Theta_3\right\| \leq \frac{1}{\widehat{\eta}}\|\widetilde{Z}^+\|_2\|Z\|_2 \left\{\left\|\begin{pmatrix} \widetilde{U}_2^H(\widetilde{A}-A)Q_{10} & \\ & \widetilde{V}_2^H(\widetilde{B}-B)Q_{10} \end{pmatrix}\right\| \|Z^+\|_2 + \frac{1}{2}\left\|\begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix}(Z-\widetilde{Z})\widetilde{Q}_{20}\right\| \|\widetilde{Z}^+\|_2 \right\},$$

(8.8)

$$\left\|\sin\Theta_4\right\| \leq \frac{1}{\widehat{\eta}}\|\widetilde{Z}\|_2\|Z^+\|_2 \left\{\frac{1}{2}\left\|\begin{pmatrix} \widetilde{U}_2^H & \\ & \widetilde{V}_2^H \end{pmatrix}(Z-\widetilde{Z})Q_{10}\right\| \|Z^+\|_2 + \left\|\begin{pmatrix} U_1^H(\widetilde{A}-A)\widetilde{Q}_{20} & \\ & V_1^H(\widetilde{B}-B)\widetilde{Q}_{20} \end{pmatrix}\right\| \|\widetilde{Z}^+\|_2 \right\},$$

where Q_{j0}, … are defined by (5.7),

$$\widehat{\eta} \overset{\text{def}}{=} \min_{\substack{(\alpha,\beta)\in\Gamma_1 \\ (\widetilde{\alpha},\widetilde{\beta})\in\Gamma_2}} \frac{|\alpha^2\widetilde{\beta}^2 - \beta^2\widetilde{\alpha}^2|}{(\alpha^2+\beta^2)(\widetilde{\alpha}^2+\widetilde{\beta}^2)}. \tag{8.9}$$

If, instead of Description I, *we use its equivalence, Description* II, *to describe the separation between $\sigma\{A_{11},B_{11}\}$ and $\sigma\{\widetilde{A}_{22},\widetilde{B}_{22}\}$, i.e., there exist $\alpha \geq 0$, $\delta > 0$, $\alpha+\delta \leq 1$ such that*

(8.10)

$$\max_{(\widetilde{\alpha}_j,\widetilde{\beta}_j)\in\sigma\{\widetilde{A}_{22},\widetilde{B}_{22}\}} \rho\big((\widetilde{\alpha}_j,\widetilde{\beta}_j),(0,1)\big) \leq \alpha, \qquad \min_{(\alpha_i,\beta_i)\in\sigma\{A_{11},B_{11}\}} \rho\big((\alpha_i,\beta_i),(0,1)\big) \geq \alpha+\delta,$$

then

$$\begin{aligned} \|\sin\Theta_3\| \leq & \frac{1}{\sqrt{1-\alpha^2}}\|\widetilde{Z}^+\|_2\|Z\|_2 \\ & \left[\left\|\widetilde{V}_2^H(\widetilde{B}-B)Q_{10}\right\| \|Z^+\|_2 + \sqrt{1-(\alpha+\delta)^2}\,\|\sin\Theta_2\|\right] \\ \leq & \|\widetilde{Z}^+\|_2\|Z\|_2 \left[\frac{1}{\eta}\left\|\widetilde{V}_2^H(\widetilde{B}-B)Q_{10}\right\| \|Z^+\|_2 + \|\sin\Theta_2\|\right], \end{aligned} \tag{8.11}$$

(8.12)

$$\begin{aligned} \|\sin\Theta_4\| \leq & \frac{1}{\alpha+\delta}\|\widetilde{Z}\|_2\|Z^+\|_2 \left[\left\|\widetilde{U}_1^H(\widetilde{A}-A)Q_{20}\right\| \|\widetilde{Z}^+\|_2 + \alpha\,\|\sin\Theta_1\|\right] \\ \leq & |\widetilde{Z}\|_2\|Z^+\|_2 \left[\frac{1}{\eta}\left\|\widetilde{U}_1^H(\widetilde{A}-A)Q_{20}\right\| \|\widetilde{Z}^+\|_2 + \|\sin\Theta_1\|\right]. \end{aligned}$$

Proof. Without loss of generality, assume that $\{A,B\}$ and $\{\widetilde{A},\widetilde{B}\}$ have decompositions as described in the beginning of §7 up to (7.3).

For the case of $m=p=n$, (8.6) follows from (5.6), (7.9), and (8.4), since by (8.1) and (8.3), we have

$$C_1 = (-\widetilde{\Omega}_2,\widetilde{\Lambda}_2)\begin{pmatrix} \underline{R}_A \\ \underline{R}_B \end{pmatrix} \Rightarrow \|C_1\| \leq \|(-\widetilde{\Omega}_2,\widetilde{\Lambda}_2)\|_2 \left\|\begin{pmatrix} \underline{R}_A \\ \underline{R}_B \end{pmatrix}\right\| = \left\|\begin{pmatrix} \underline{R}_A \\ \underline{R}_B \end{pmatrix}\right\|,$$

$$C_2^H = (\Omega_1,-\Lambda_1)\begin{pmatrix} \underline{R}_{A^H}^H \\ \underline{R}_{B^H}^H \end{pmatrix} \Rightarrow \|C_2\| \leq \|(\Omega_1,-\Lambda_1)\|_2 \left\|\begin{pmatrix} \underline{R}_{A^H}^H \\ \underline{R}_{B^H}^H \end{pmatrix}\right\| = \left\|\begin{pmatrix} \underline{R}_{A^H}^H \\ \underline{R}_{B^H}^H \end{pmatrix}\right\|,$$

and

(8.13)
$$\left\|\begin{pmatrix}\underline{R}_A\\ \underline{R}_B\end{pmatrix}\right\| = \left\|\begin{pmatrix}\widetilde{U}_2^H & \\ & \widetilde{V}_2^H\end{pmatrix}(Z-\widetilde{Z})Q_1\right\| \le \left\|\begin{pmatrix}\widetilde{U}_2^H & \\ & \widetilde{V}_2^H\end{pmatrix}(Z-\widetilde{Z})Q_{10}\right\| \|Z^+\|_2,$$

(8.14)
$$\left\|\begin{pmatrix}\underline{R}_{A^H}^H\\ \underline{R}_{B^H}^H\end{pmatrix}\right\| = \left\|\begin{pmatrix}U_1^H & \\ & V_1^H\end{pmatrix}(Z-\widetilde{Z})\widetilde{Q}_2\right\| \le \left\|\begin{pmatrix}U_1^H & \\ & V_1^H\end{pmatrix}(Z-\widetilde{Z})\widetilde{Q}_{20}\right\| \|\widetilde{Z}^+\|_2.$$

Here we have used the following facts in the above derivation.

$$\|(-\widetilde{\Omega}_2, \widetilde{\Lambda}_2)\|_2 = \max_j \sqrt{\widetilde{\alpha}_j^2 + \widetilde{\beta}_j^2} = 1, \qquad \|(\Omega_1, -\Lambda_1)\|_2 = \max_i \sqrt{\alpha_i^2+\beta_i^2} = 1,$$
$$Q_1 = Q_1(Q_1^HQ_1)^{-\frac{1}{2}}(Q_1^HQ_1)^{\frac{1}{2}} = Q_{10}(Q_1^HQ_1)^{\frac{1}{2}}, \qquad \widetilde{Q}_2 = \widetilde{Q}_{20}(\widetilde{Q}_2^H\widetilde{Q}_2)^{\frac{1}{2}},$$
$$\|(Q_1^HQ_1)^{\frac{1}{2}}\|_2 = \|Q_1\|_2 \le \|(Q_1,Q_2)\|_2 \le \|Z^+\|_2 \quad \text{(by (1.6))},$$
$$\|(\widetilde{Q}_2^H\widetilde{Q}_2)^{\frac{1}{2}}\|_2 \le \|\widetilde{Z}^+\|_2.$$

Inequalities (8.7) and (8.8) follow from the proof of (7.9) in §7 (see also Li [18]), (8.5), (8.13), (8.14), Lemmas 1.1 and 1.2, and (5.6), since

$$\begin{aligned} &\left\|\widetilde{P}_2^HQ_1\right\| \ge \|(\widetilde{P}_2^H\widetilde{P}_2)^{-\frac{1}{2}}\|_2^{-1}\left\|\widetilde{P}_{20}^HQ_{10}\right\| \|(Q_1^HQ_1)^{-\frac{1}{2}}\|_2^{-1} \\ &\qquad \ge \|\widetilde{P}^{-1}\|_2^{-1}\|\sin\Theta_3\|\|Q^{-1}\|_2^{-1}, \end{aligned} \tag{8.15}$$

$$\Rightarrow \|\sin\Theta_3\| \le \|Z\|_2\|\widetilde{Z}^+\|_2\left\|\widetilde{P}_2^HQ_1\right\|, \tag{8.16}$$

$$\|\sin\Theta_4\| \le \|Z^+\|_2\|\widetilde{Z}\|_2\left\|\widetilde{Q}_2^HP_1\right\|; \tag{8.17}$$

and

$$\begin{aligned} C_5 &= (\widetilde{\Lambda}_2\underline{R}_A + \underline{R}_{A^H}\Lambda_1)\Omega_1^2 - (\widetilde{\Omega}_2\underline{R}_B + \underline{R}_{B^H}\Omega_1)\Lambda_1^2 \\ &= (\widetilde{\Lambda}_2, -\widetilde{\Omega}_2)\begin{pmatrix}\underline{R}_A & \\ & \underline{R}_B\end{pmatrix}\begin{pmatrix}\Omega_1^2\\ \Lambda_1^2\end{pmatrix} + (\underline{R}_{A^H}, \underline{R}_{B^H})\begin{pmatrix}\Omega_1\\ -\Lambda_1\end{pmatrix}\Omega_1\Lambda_1, \\ C_6 &= -\widetilde{\Omega}_2^2(\widetilde{\Lambda}_2\underline{R}_A + \underline{R}_{A^H}\Lambda_1) + \widetilde{\Lambda}_2^2(\widetilde{\Omega}_2\underline{R}_B + \underline{R}_{B^H}\Omega_1) \\ &= \widetilde{\Lambda}_2\widetilde{\Omega}_2(-\widetilde{\Omega}_2, \widetilde{\Lambda}_2)\begin{pmatrix}\underline{R}_A\\ \underline{R}_B\end{pmatrix} + (\widetilde{\Omega}_2^2, \widetilde{\Lambda}_2^2)\begin{pmatrix}\underline{R}_{A^H} & \\ & \underline{R}_{B^H}\end{pmatrix}\begin{pmatrix}-\Lambda_1\\ \Omega_1\end{pmatrix}, \end{aligned}$$

which produce

$$\begin{aligned} \|C_5\| &\le \left\|\begin{pmatrix}\underline{R}_A & \\ & \underline{R}_B\end{pmatrix}\right\|\left\|\begin{pmatrix}\Omega_1^2\\ \Lambda_1^2\end{pmatrix}\right\|_2 + \|(\underline{R}_{A^H}, \underline{R}_{B^H})\|\,\|\Omega_1\Lambda_1\|_2 \\ &\le \left\|\begin{pmatrix}\underline{R}_A & \\ & \underline{R}_B\end{pmatrix}\right\| + \frac{1}{2}\left\|\begin{pmatrix}\underline{R}_{A^H}^H\\ \underline{R}_{B^H}^H\end{pmatrix}\right\|, \\ \|C_6\| &\le \frac{1}{2}\left\|\begin{pmatrix}\underline{R}_A\\ \underline{R}_B\end{pmatrix}\right\| + \left\|\begin{pmatrix}\underline{R}_{A^H}^H & \\ & \underline{R}_{B^H}^H\end{pmatrix}\right\|; \end{aligned} \tag{8.18}$$

and

$$\left\|\begin{pmatrix} \underline{R}_A & \\ & \underline{R}_B \end{pmatrix}\right\| = \left\|\begin{pmatrix} \widetilde{U}_2^H(\widetilde{A}-A)Q_{10} & \\ & \widetilde{V}_2^H(\widetilde{B}-B)Q_{10} \end{pmatrix}\right\| \|Z^+\|_2,$$
$$\left\|\begin{pmatrix} \underline{R}_{A^H}^H & \\ & \underline{R}_{B^H}^H \end{pmatrix}\right\| = \left\|\begin{pmatrix} U_1^H(\widetilde{A}-A)\widetilde{Q}_{20} & \\ & V_1^H(\widetilde{B}-B)\widetilde{Q}_{20} \end{pmatrix}\right\| \|\widetilde{Z}^+\|_2.$$

In the derivation of (8.18), we have noted

$$\|\Omega_1\Lambda_1\|_2 = \max_i \alpha_i\beta_i \le \frac{1}{2}, \qquad \left\|\begin{pmatrix} \Omega_1^2 \\ \Lambda_1^2 \end{pmatrix}\right\|_2 = \max_i \sqrt{\alpha_i^4+\beta_i^4} \le 1,$$

and some other inequalities of the same kinds. In proving the last relation in (8.15), we need

$$\|(Q_1^HQ_1)^{-\frac{1}{2}}\|_2^{-1} = \sigma_{\min}(Q_1) \ge \sigma_{\min}(Q),$$

where $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix. Inequalities (8.11) and (8.12) follow from the second and the third equations in (8.3), (8.16), (8.17), and (8.10), since $\alpha+\delta \ge \eta$, $\sqrt{1-\alpha^2} \ge \eta$ by (5.11), and

$$\|\Lambda_1^{-1}\|_2^{-1} \ge \alpha+\delta, \quad \|\Omega_1\|_2 \le \sqrt{1-(\alpha+\delta)^2}, \quad \|\widetilde{\Lambda}_2\|_2 \le \alpha, \quad \|\widetilde{\Omega}_2^{-1}\|_2^{-1} \ge \sqrt{1-\alpha^2},$$
$$\sqrt{1-\alpha^2}\left\|\widetilde{P}_2^HQ_1\right\| \le \|\widetilde{\Omega}_2^{-1}\|_2^{-1}\left\|\widetilde{P}_2^HQ_1\right\| \le \left\|\widetilde{\Omega}_2\widetilde{P}_2^HQ_1\right\|$$
$$\le \|\underline{R}_B\| + \left\|\widetilde{V}_2^HV_1\right\|\sqrt{1-(\alpha+\delta)^2},$$
$$(\alpha+\delta)\left\|\widetilde{Q}_2^HP_1\right\| \le \|\underline{R}_{A^H}\| + \alpha\left\|\widetilde{U}_2^HU_1\right\|.$$

The proof of the theorem for the case of $m \ne n$ or $p \ne n$ can be given in an analogous way to those in §7. We omit the details here. □

The following theorem, in a somewhat similar spirit to Theorems 6.2 and 6.3, can also be proved analogously.

THEOREM 8.2. *The conditions and notations are as described in Theorem 6.2. Then there exists a constant $c \ge 1$ such that for any unitarily invariant norm $\|\cdot\|$ we have*

$$\left\|\begin{pmatrix} \sin\Theta_1 & \\ & \sin\Theta_2 \end{pmatrix}\right\| \le c\cdot\frac{1}{\tilde{\tilde{\eta}}}\left[\left\|\begin{pmatrix} \widetilde{U}_2^H & \\ & \widetilde{V}_2^H \end{pmatrix}(Z-\widetilde{Z})Q_{10}\right\| \|Z^+\|_2 + \left\|\begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix}(Z-\widetilde{Z})\widetilde{Q}_{20}\right\| \|\widetilde{Z}^+\|_2\right], \tag{8.19}$$

and

(8.20)
$$\|\sin\Theta_3\| \le c\cdot\frac{1}{\tilde{\tilde{\hat{\eta}}}}\|\widetilde{Z}^+\|_2\|Z\|_2\left[\left\|\begin{pmatrix} \widetilde{U}_2^H(\widetilde{A}-A)Q_{10} & \\ & \widetilde{V}_2^H(\widetilde{B}-B)Q_{10} \end{pmatrix}\right\| \|Z^+\|_2 + \frac{1}{2}\left\|\begin{pmatrix} U_1^H & \\ & V_1^H \end{pmatrix}(Z-\widetilde{Z})\widetilde{Q}_{20}\right\| \|\widetilde{Z}^+\|_2\right],$$

(8.21)
$$\|\sin\Theta_4\| \le c\cdot\frac{1}{\tilde{\tilde{\hat{\eta}}}}\|\widetilde{Z}\|_2\|Z^+\|_2\left[\frac{1}{2}\left\|\begin{pmatrix} \widetilde{U}_2^H & \\ & \widetilde{V}_2^H \end{pmatrix}(Z-\widetilde{Z})Q_{10}\right\| \|Z^+\|_2 + \left\|\begin{pmatrix} U_1^H(\widetilde{A}-A)\widetilde{Q}_{20} & \\ & V_1^H(\widetilde{B}-B)\widetilde{Q}_{20} \end{pmatrix}\right\| \|\widetilde{Z}^+\|_2\right],$$

where

$$
\text{(8.22)}\qquad
\begin{aligned}
\underline{\widehat{\eta}} &\overset{\mathrm{def}}{=} \min_{\substack{1\le i\le \ell\\ \ell+1\le j\le n}} \frac{|\alpha_i^2\widetilde{\beta}_j^2-\beta_i^2\widetilde{\alpha}_j^2|}{(\alpha_i^2+\beta_i^2)(\widetilde{\alpha}_j^2+\widetilde{\beta}_j^2)},\\
\widehat{\eta}_1 &\overset{\mathrm{def}}{=} \min_{\ell+1\le j\le n} \frac{|1^2\widetilde{\beta}_j^2-0^2\widetilde{\alpha}_j^2|}{(1^2+0^2)(\widetilde{\alpha}_j^2+\widetilde{\beta}_j^2)},\\
\widehat{\eta}_2 &\overset{\mathrm{def}}{=} \min_{1\le i\le \ell} \frac{|\alpha_i^2 0^2-\beta_i^2 1^2|}{(\alpha_i^2+\beta_i^2)(0^2+1^2)},
\end{aligned}
$$

and

$$
\text{(8.23)}\qquad 0<\widetilde{\widehat{\eta}} \overset{\mathrm{def}}{=} \begin{cases}
\underline{\widehat{\eta}} & \textit{if } m\le n \quad \textit{and} \quad p\le n,\\
\min\{\underline{\widehat{\eta}},\widehat{\eta}_1\} & \textit{if } m\le n<p,\\
\min\{\underline{\widehat{\eta}},\widehat{\eta}_2\} & \textit{if } m>n\ge p,\\
\min\{\underline{\widehat{\eta}},\widehat{\eta}_1,\widehat{\eta}_2\} & \textit{if } m>n \quad \textit{and} \quad p>n.
\end{cases}
$$

Generally, c can be chosen so that (6.10) *holds, but for* $\|\cdot\|=\|\cdot\|_F$, (8.19)–(8.21) *can be improved in the following way: replacing c in* (8.20) *and* (8.21) *by* 1 *and*

(8.24)

$$
\begin{aligned}
[\|\sin\Theta_1\|_F^2+\|\sin\Theta_2\|_F^2]^{\frac{1}{2}} \le \frac{1}{\widetilde{\widehat{\eta}}}\Bigg[&\left\|\begin{pmatrix}\widetilde{U}_2^H & \\ & \widetilde{V}_2^H\end{pmatrix}(Z-\widetilde{Z})Q_{10}\right\|_F^2 \|Z^+\|_2^2\\
&+\left\|\begin{pmatrix}U_1^H & \\ & V_1^H\end{pmatrix}(Z-\widetilde{Z})\widetilde{Q}_{20}\right\|_F^2 \|\widetilde{Z}^+\|_2^2\Bigg]^{\frac{1}{2}}.
\end{aligned}
$$

We have illustrated in detail in §6 how much weaker the gap $\widetilde{\eta}$ (defined by (6.5)) is than η (defined by (5.9)). Similar arguments hold for the weakness of $\widetilde{\widehat{\eta}}$ defined by (8.23) with respect to $\widehat{\eta}$ defined by (8.9). There is a straightforward interpretation of the appearance of $\widehat{\eta}$ (cf. $\widetilde{\widehat{\eta}}$). We now have an idea of how the generalized singular value problem closely relates to the generalized eigenvalue problem for definite pencils. Specifically, $(\alpha,\beta)\in\sigma\{A,B\}\Rightarrow(\alpha^2,\beta^2)\in\lambda(A^HA,B^HB)$, the subspace spanned by the column vectors of Q_1, is an eigenspace of the definite pencil $A^HA-\lambda B^HB$ belonging to the generalized eigenvalues (α_i^2,β_i^2), $i=1,2,\ldots,\ell$, and so on. $\widehat{\eta}$ (cf. $\widetilde{\widehat{\eta}}$) should then be regarded as a gap in a sense of the chordal metric between two certain subsets of the spectra of the corresponding definite pencils, since for α, β, $\widetilde{\alpha}$, $\widetilde{\beta}\ge 0$,

$$
\frac{1}{\sqrt{2}}\rho\big((\alpha^2,\beta^2),(\widetilde{\alpha}^2,\widetilde{\beta}^2)\big)\le\frac{|\alpha^2\widetilde{\beta}^2-\beta^2\widetilde{\alpha}^2|}{(\alpha^2+\beta^2)(\widetilde{\alpha}^2+\widetilde{\beta}^2)}\le\rho\big((\alpha^2,\beta^2),(\widetilde{\alpha}^2,\widetilde{\beta}^2)\big).
$$

In this case, bounds (8.7), (8.8), (8.20), and (8.21) reflect on the hidden relation between the problems, though it might be impossible to get these bounds via those for the eigenspaces for definite pencils [17], [18], [32].

As to the sharpness of our bounds in §§6 and 8, to our knowledge, it may be impossible to know exactly which one is better than others. But generally we propose that, apart from weak assumptions on gaps between certain sets of GSVs, bounds

(6.2), (6.3), (6.6), (6.7), (6.8), (6.9), (8.6), and (8.19) are equally effective. They may be only a constant factor apart. For $\sin\Theta_3$ and $\sin\Theta_4$, bounds (8.7) and (8.8) may be much less sharp than (8.11) and (8.12). To see this, we note

$$\eta \geq \widehat{\eta} \geq \eta^2,$$

and $\widehat{\eta}$ approaching η^2 indeed occurs in some cases. Thus, if η is also very small, then the right-hand sides of (8.7) and (8.8) must be much larger than those of (8.11) and (8.12). Why should we bother to establish (8.7) and (8.8)? We argue that (8.7) and (8.8) are of theoretical interest. Also, the perturbation equations (8.5) used to prove (8.7) and (8.8) can be used to prove (8.20) and (8.21), which are valid under weaker assumptions that certain subsets of GSVs of corresponding matrix pairs are not required to be well separated, as shown by Fig. 2 and (6.1).

As our derivation for bounds seems a little more elegant than that of Sun [30], it would be natural for us to expect that our bounds should be better. To enable us to compare ours with Sun's, we first simplify bounds due to Sun [30]. In our process of simplification, no essential amplification is involved, i.e., there is no essential loss of sharpness in the simplified bounds in comparison with the original ones of Sun [30]. We note that Sun [30] presented only bounds for $\sin\Theta_i$, $i = 1, 2, 3, 4$, under the assumptions of Theorem 6.1. So now we assume that the conditions and notations are as described in Theorem 6.1, and Description II (refer to (8.10)), as a description equivalent to Description I, is also referred to when suitable; η is defined by (5.11). In our notation, Sun's bounds are read as

(8.25)
$$\|\sin\Theta_1\| \leq \frac{\tau_1}{\delta}\left[(r_a + \omega_1 r_b)\|Z^+\|_2 + \omega_1\omega_2(\omega_3 r_a' + r_b')\|\widetilde{Z}^+\|_2\right],$$

(8.26)
$$\|\sin\Theta_2\| \leq \frac{\tau_2}{\delta}\left[\omega_3\omega_4(r_a + \omega_1 r_b)\|Z^+\|_2 + (\omega_3 r_a' + r_b')\|\widetilde{Z}^+\|_2\right],$$

(8.27)
$$\|\sin\Theta_3\| \leq \frac{\tau_2}{\delta}\|Z\|_2\|\widetilde{Z}^+\|_2\left[(\omega_2\omega_3\omega_4 r_a + r_b)\|Z^+\|_2 + \omega_2(\omega_3 r_a' + r_b')\|\widetilde{Z}^+\|_2\right],$$

(8.28)
$$\|\sin\Theta_4\| \leq \frac{\tau_1}{\delta}\|Z^+\|_2\|\widetilde{Z}\|_2\left[\omega_4(r_a + \omega_1 r_b)\|Z^+\|_2 + (r_a' + \omega_1\omega_2\omega_4 r_b')\|\widetilde{Z}^+\|_2\right],$$

where

$$\tau_1 = \frac{(\alpha+\delta)(1-\alpha^2)}{2\alpha+\delta}, \qquad \tau_2 = \frac{(\alpha+\delta)^2\sqrt{1-\alpha^2}}{2\alpha+\delta},$$
$$\omega_1 = \frac{\alpha}{\sqrt{1-\alpha^2}}, \quad \omega_2 = \sqrt{\frac{1-(\alpha+\delta)^2}{1-\alpha^2}}, \quad \omega_2 = \frac{\sqrt{1-(\alpha+\delta)^2}}{\alpha+\delta}, \quad \omega_4 = \frac{\alpha}{\alpha+\delta},$$
$$r_a = \left\|(\widetilde{A}-A)Q_{10}\right\|, r_b = \left\|(\widetilde{B}-B)Q_{10}\right\|, r_a' = \left\|U_1^H(\widetilde{A}-A)\right\|, r_b' = \left\|V_1^H(\widetilde{B}-B)\right\|.$$

Denote $d = (\alpha+\delta)\sqrt{1-\alpha^2} + \alpha\sqrt{1-(\alpha+\delta)^2}$. For $\sin\Theta_1$ and $\sin\Theta_2$, it follows from

$$\eta\frac{\tau_1}{\delta} = \frac{(\alpha+\delta)\sqrt{1-\alpha^2}}{d}\sqrt{1-\alpha^2}, \qquad \eta\frac{\tau_1}{\delta}\omega_1 = \frac{(\alpha+\delta)\sqrt{1-\alpha^2}}{d}\alpha,$$
$$\eta\frac{\tau_1}{\delta}\omega_1\omega_2 = \frac{\alpha\sqrt{1-(\alpha+\delta)^2}}{d}(\alpha+\delta), \quad \eta\frac{\tau_1}{\delta}\omega_1\omega_2\omega_3 = \frac{\alpha\sqrt{1-(\alpha+\delta)^2}}{d}\sqrt{1-(\alpha+\delta)^2},$$

and

$$\eta\frac{\tau_2}{\delta} = \frac{(\alpha+\delta)\sqrt{1-\alpha^2}}{d}(\alpha+\delta), \qquad \eta\frac{\tau_2}{\delta}\omega_3 = \frac{(\alpha+\delta)\sqrt{1-\alpha^2}}{d}\sqrt{1-(\alpha+\delta)^2},$$
$$\eta\frac{\tau_2}{\delta}\omega_3\omega_4 = \frac{\alpha\sqrt{1-(\alpha+\delta)^2}}{d}\sqrt{1-\alpha^2}, \qquad \eta\frac{\tau_2}{\delta}\omega_3\omega_4\omega_1 = \frac{\alpha\sqrt{1-(\alpha+\delta)^2}}{d}\alpha$$

that

$$\max_{i=1,2}\|\sin\Theta_i\| \le \frac{1}{\eta}\max\left\{\sqrt{r_a^2+r_b^2}\,\|Z^+\|_2, \sqrt{{r'_a}^2+{r'_b}^2}\,\|\widetilde{Z}^+\|_2\right\}. \tag{8.29}$$

The simplifications of (8.27) and (8.28) are in the same spirit. Here we give the final results and leave the details to the reader:

$$\|\sin\Theta_3\| \le \frac{1}{\eta}\cdot 2\|Z\|_2\|\widetilde{Z}^+\|_2\max\left\{\sqrt{r_a^2+r_b^2}\,\|Z^+\|_2, \sqrt{{r'_a}^2+{r'_b}^2}\,\|\widetilde{Z}^+\|_2\right\}, \tag{8.30}$$

$$\|\sin\Theta_4\| \le \frac{1}{\eta}\cdot 2\|\widetilde{Z}\|_2\|Z^+\|_2\max\left\{\sqrt{r_a^2+r_b^2}\,\|Z^+\|_2, \sqrt{{r'_a}^2+{r'_b}^2}\,\|\widetilde{Z}^+\|_2\right\}. \tag{8.31}$$

Now turning to our bounds (8.6), (8.11), and (8.12), we note that

$$\begin{aligned}
\left\|\begin{pmatrix}\widetilde{U}_2^H & \\ & \widetilde{V}_2^H\end{pmatrix}(Z-\widetilde{Z})Q_{10}\right\| &\le \left\|\begin{pmatrix}A-\widetilde{A}\\ B-\widetilde{B}\end{pmatrix}Q_{10}\right\| \\
&\le \sqrt{2}\sqrt{r_a^2+r_b^2} \le 2\left\|\begin{pmatrix}A-\widetilde{A}\\ B-\widetilde{B}\end{pmatrix}Q_{10}\right\|, \\
\left\|\begin{pmatrix}U_1^H & \\ & V_1^H\end{pmatrix}(Z-\widetilde{Z})\widetilde{Q}_{20}\right\| &\le \left\|\begin{pmatrix}U_1^H & \\ & V_1^H\end{pmatrix}\begin{pmatrix}A-\widetilde{A}\\ B-\widetilde{B}\end{pmatrix}\right\| \\
&\le \sqrt{2}\sqrt{{r'_a}^2+{r'_b}^2} \le 2\left\|\begin{pmatrix}U_1^H & \\ & V_1^H\end{pmatrix}\begin{pmatrix}A-\widetilde{A}\\ B-\widetilde{B}\end{pmatrix}\right\|.
\end{aligned}$$

Therefore (8.6) and (8.29), thus (8.25) and (8.26) as well, would be equally effective. By the way, we should say that the simplified bounds (8.29)–(8.31) of Sun's bounds seem more favorable than their original (8.25)–(8.28).

Acknowledgment. The author acknowledges helpful advice from the referees. He also acknowledges that an anonymous referee pointed out that bounds for $\sin\Theta_3$ are of great practical interest and suggested the inclusion of such bounds in the revised version of this paper. He feels indebted to Professor Axel Ruhe for reading the manuscript so carefully and for very detailed and constructive suggestions for revising this paper.

REFERENCES

[1] R. Bhatia, C. Davis, and A. McIntosh, *Perturbation of spectral subspaces and the solution of linear equations*, Linear Algebra Appl., 52/53 (1983), pp. 45–67.

[2] R. Bhatia and C. Davis, *A bound for the spectral variation of a unitary operator*, Linear Multilinear Algebra, 15 (1984), pp. 71–76.

[3] R. Bhatia, C. Davis, and P. Koosis, *An extremal problem in Fourier analysis with applications to operator theory*, J. Funct. Anal., 82 (1989), pp. 138–150.

[4] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.

[5] L. Eldén, *A weighted pseudoinverse, generalized singular values, and constrained least squares problems*, BIT, 22 (1982), pp. 487–502.

[6] I. C. Gohberg and M. G. Kreĭn, *Introduction to the Theory of Linear Nonselfadjoint Operators,* Trans. Math. Monographs, 18, American Mathematical Society, Providence, RI, 1969.

[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[8] G. H. Golub and J. H. Wilkinson, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Rev., 18 (1976), pp. 578–619.

[9] P. C. Hansen, *The truncated* SVD *as a method for regularization*, BIT, 27 (1987), pp. 534–553.

[10] ———, *Regularization,* GSVD *and truncated* GSVD, BIT, 29 (1989), pp. 491–504.

[11] A. J. Hoffman and H. W. Wielandt, *The variation of the spectrum of a normal matrix*, Duke Math. J., 20 (1953), pp. 37–39.

[12] B. Kågström, *The generalized singular value decomposition and the general $(A-\lambda B)$-problem*, BIT, 24 (1984), pp. 568–583.

[13] ———, RGSVD—*an algorithm for computing the Kronecker structure and reducing subspaces of singular $A-\lambda B$ pencils*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 185–211.

[14] B. Kågström and A. Ruhe, *Matrix Pencils,* Proc. Pite Havsbad, Lecture Notes in Mathematics 973, Springer-Verlag, Berlin, 1983.

[15] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[16] R.-C. Li, *On the variations of the spectra of matrix pencils*, Linear Algebra Appl., 139 (1990), pp. 147–164.

[17] ———, *On perturbations of matrix pencils with real spectra,*, Preprint, Institute of Applied Physics and Computational Mathematics, Beijing, 1989; Math. Comp., to appear.

[18] ———, *Solution of linear matrix equation $AXD-BXC=S$ and perturbation of eigenspaces of a matrix pencil*, J. Comput. Math., to appear.

[19] ———, *Norms of certain matrices with applications to variations of the spectra of matrices and matrix pencils*, Linear Algebra Appl., 1992, to appear.

[20] ———, *Spectral variations and Hadamard products*, Preprint, Computing Center, Academia Sinica, Beijing, China, 1988.

[21] V. B. Lidskii, *The proper values of the sum and product of symmetric matrices*, Dokl. Akad. Nauk SSSR, 75 (1950), pp. 769–772.

[22] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[23] L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, Quart. J. Math. Oxford, 11 (1960), pp. 50–59.

[24] C. C. Paige, *A note on a result of Sun Ji-guang: Sensitivity of the* CS *and* GSV *decompositions*, SIAM J. Numer. Anal., 21 (1984), pp. 186–191.

[25] C. C. Paige and M. A. Saunders, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.

[26] B. Schwarz and A. Zaks, *Geometries of the projective matrix spaces*, J. Algebra, 95 (1985), pp. 264–307.

[27] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[28] ———, *Perturbation bounds for the definite generalized eigenvalue problem*, Linear Algebra Appl., 23 (1979), pp. 69–83.

[29] J.-G. Sun, *The perturbation bounds of generalized eigenvalues of a class of matrix-pairs*, Math. Numer. Sinica, 4 (1982), pp. 23–29. (In Chinese.)

[30] ———, *Perturbation analysis for the generalized singular value problem*, SIAM J. Numer. Anal., 20 (1983), pp. 611–625.

[31] ———, *A note on Stewart's theorem for definite matrix pairs*, Linear Algebra Appl., 48 (1982), pp. 331–339.

[32] ———, *The perturbation bounds for eigenspaces of a definite matrix-pair*, Numer. Math., 41 (1983), pp. 321–343.

[33] J.-G. Sun, *On the perturbation of generalized singular values*, Math. Numer. Sinica, 4 (1982), pp. 229–233. (In Chinese.)

[34] J.-G. Sun and G. W. Stewart, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[35] J.-G. Sun, *Perturbation theorems for generalized singular values*, J. Comput. Math., 1 (1983), pp. 233–242.

[36] P. Van Dooren, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–140.

[37] C. F. Van Loan, *A general matrix eigenvalue algorithm*, SIAM J. Numer. Anal., 12 (1975), pp. 819–834.

[38] ———, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.

[39] J. Von Neumann, *Some matrix-inequalities and metrization of matrix-space*, Bull. Inst. Math. Mécan., Univ. Kouybycheff Tomsk, 1 (1935–1937), pp. 286–300.

[40] P.-Å. Wedin, *Perturbation bounds in connection with singular value decomposition*, BIT, 12 (1972), pp. 99–111.

[41] H. Weyl, *Das asymptotische Verteilungsgesetz der Eigenwerte linear partielles Differentialgleichungen*, Math. Ann., 71 (1912), pp. 441–479.

[42] H. Wielandt, *An extremum property of sums of eigenvalues*, Proc. Amer. Math. Soc., 6 (1955), pp. 106–110.

[43] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 235–241, January 1993

018

PRECISE MATRIX EIGENVALUES USING RANGE ARITHMETIC*

OLIVER ABERTH[†] AND MARK J. SCHAEFER[‡]

Abstract. A recent paper by Aberth and Schaefer [*Trans. Math. Software*, to appear] described how the programming language C++ can serve as a vehicle for replacing floating-point arithmetic with the more advanced range arithmetic. This arithmetic makes it easier to generate accurate answers to scientific computation problems. This paper discusses the use of range arithmetic to solve precisely the eigenvalue problem for square matrices of modest size, that is, of about twenty rows or less.

Key words. matrix eigenvalues, range arithmetic, C++ language, precise computation

AMS(MOS) subject classification. 65F15

1. Introduction.

Range arithmetic is a variable precision, interval arithmetic system. Here each number has the usual fields of ordinary floating-point arithmetic, that is, a sign, an exponent, and a mantissa, but the mantissa is permitted to vary in length. A global constant under program control called *precision* sets an upper bound on the mantissa length of numbers produced by computation. Each number has an additional small field, called the range, giving the number's maximum error or interval width. This field is zero for input constants, since these are without error, but quickly becomes nonzero for computed numbers derived from constants. The range field is calculated according to interval concepts [4], [6], but its small size keeps the speed of the arithmetic comparable to that of ordinary floating-point arithmetic. As a long computation proceeds, the range field causes the gradual discard of suspect trailing mantissa digits, so that mantissas tend to get shorter. This effect depends also on the precision setting, with a higher setting leading to more mantissa digits being retained. A program can monitor the mantissas of key results and adjust the precision accordingly.

As an example consider the Gaussian elimination method of solving a system of nonsingular linear equations. With range arithmetic an initial precision is chosen, more or less by guessing, but the choice can reflect the number of correct decimal places desired and the number of equations to be solved. After the elimination and back substitution procedures are carried out, the answers are checked to determine whether the required number of decimal places have been obtained. If not, the discrepancy between the number of correct decimals obtained and the number required is used to estimate the higher precision needed, and after the precision is reset, the computation routine is re-entered. With a good choice of the initial precision, the first computation pass usually succeeds. Cases where the computation must be repeated are likely to be cases where ordinary floating-point arithmetic would yield poor results.

For a square matrix A, finding its eigenvalues precisely with range arithmetic is somewhat more complicated. We consider here methods that solve this problem by transforming A to upper triangular form $B = T^{-1}AT$ so that the diagonal entries

*Received by the editors December 26, 1990; accepted for publication (in revised form) July 5, 1991. This work was supported by National Science Foundation grant CCR-8900489.

[†]Department of Mathematics, Texas A & M University, College Station, Texas 77843 (aberth@math.tamu.edu).

[‡]Universität Tübingen, Wilhelm-Schickhard-Institut, Sand 13, 7400 Tübingen, Germany.

of B yield the eigenvalues. The range of a diagonal element of B does not give its complete error as an eigenvalue because B is usually only in "near upper triangular form," that is, the elements of B below the diagonal are small but not necessarily zero. An additional computation is required to bound the error in the eigenvalues caused by the impure form of B. After eigenvalues have their ranges increased to reflect this additional error, the test for the required number of decimal places and a recycling of the procedure where necessary are as described before.

In the next section we present a method for bounding the eigenvalue errors of near upper triangular matrices that is based on the well-known Gershgorin Theorem and generalizes an approach of Wilkinson [8, pp. 638–639]. We made use of these bounds to develop a program for obtaining precise eigenvalues, first using the QR method for achieving the near upper triangular form. Unfortunately, for reasons given in §3 where the procedure is described, the program proved disappointingly slow in delivering answers. A second, more successful, eigenvalue program utilizing an adaptation of the Danilevsky method is described in the succeeding section.

2. Eigenvalue error bounds for near upper triangular matrices.

For reference, we list the basis for our eigenvalue error estimation.

GERSHGORIN'S THEOREM. *Let B be an n-square matrix with real or complex elements b_{ij}. In the complex plane define the n disks*

$$|z - b_{ii}| \leq \sum_{\substack{j=1\\ j\neq i}}^{n} |b_{ij}|, \qquad i = 1, 2, \ldots, n.$$

These n disks may or may not overlap; in general, they form a number of disjoint connected sets. Every eigenvalue of B lies in one of these connected sets, and, counting multiplicities, each connected set consisting of k disks contains k eigenvalues.

For a complex n-square matrix B define the two quantities

$$\operatorname{upper}(B) = \max_i \sum_{j=i+1}^{n} |b_{ij}| \qquad \text{and} \qquad \operatorname{lower}(B) = \max_i \sum_{j=1}^{i-1} |b_{ij}|.$$

That is, $\operatorname{upper}(B)$ is the maximum row sum of element absolute values for elements above the diagonal, and $\operatorname{lower}(B)$ is similarly defined for elements below the diagonal. For a near triangular matrix B, $\operatorname{upper}(B)$ may be sizeable but $\operatorname{lower}(B)$ is small. Assume that δ is an eigenvalue error bound we are trying to achieve. We attempt to ensure that every eigenvalue lies in one of the δ-disks $|z - b_{ii}| \leq \delta$, $i = 1, 2, \ldots, n$. As with the Gershgorin circles, these disks may overlap and, in general, form a certain number of connected sets.

THEOREM 1. *Let the n-square matrix B have δ-disks which form a number of connected sets, with at most t disks to a set. Let d be the minimum distance between any two connected sets and take d equal to upper(B) if there is only one set. If $q = \min\{1, d/\operatorname{upper}(B)\}$ and $r = \min\{1, \delta/(2\operatorname{upper}(B)\, q)\}$, then if lower$(B) \leq \frac{\delta}{2} r^t q^{n-1}$ every eigenvalue of B lies in a connected set, and counting multiplicities, every connected set of k disks contains k eigenvalues.*

To prove this result let Q be the n-square diagonal matrix with diagonal elements $q^{n-1}, q^{n-2}, \ldots, q, 1$. The matrix $B' = QBQ^{-1}$ has the same diagonal elements as B, and being a transform of B, has the same eigenvalues. It satisfies the relation

$$\operatorname{upper}(B') \leq \operatorname{upper}(B)\, q \leq d,$$

since we have

$$\sum_{j=i+1}^{n} |b'_{ij}| = \sum_{j=i+1}^{n} |b_{ij}| q^{j-i} \leq \left(\sum_{j=i+1}^{n} |b_{ij}| \right) q \leq \text{upper}\,(B)\, q.$$

For each connected set formed by the δ-disks with centers $b_{i_1 i_1}, b_{i_2 i_2}, \ldots, b_{i_k i_k}$, we define a further transformation from B' to B'' which again leaves diagonal elements and eigenvalues unchanged. Define the n-square diagonal matrix R with diagonal elements equal to r^k for rows 1 to i_1, equal to r^{k-1} for rows $i_1 + 1$ to $i_2, \ldots,$ equal to r for rows $i_{k-1} + 1$ to i_k, and equal to 1 for all rows following i_k. If we set $B'' = RB'R^{-1}$, then for rows $i_1, i_2, \ldots,$ or i_k, the sum of absolute values of elements above the diagonal is

$$\sum_{j=i+1}^{n} |b''_{ij}| = \sum_{j=i+1}^{n} |b'_{ij}| r^{m_{ij}} \leq \sum_{j=i+1}^{n} |b'_{ij}| r \leq \text{upper}\,(B')\, r \leq \text{upper}\,(B)\, qr \leq \frac{\delta}{2},$$

since for these rows an exponent m_{ij} is never less than 1. The corresponding sum for the other rows is

$$\sum_{j=i+1}^{n} |b''_{ij}| = \sum_{j=i+1}^{n} |b'_{ij}| r^{m_{ij}} \leq \sum_{j=i+1}^{n} |b'_{ij}| \leq \text{upper}\,(B') \leq d,$$

since here an exponent m_{ij} is never negative. Each element b''_{ij} below the diagonal has magnitude not greater than $|b_{ij}|/r^t q^{n-1}$. Therefore, any row sum of absolute values of these elements cannot exceed lower $(B)/r^t q^{n-1} \leq \frac{\delta}{2}$. Thus the Gershgorin radius for a diagonal element of B'' with row index chosen from $i_1, i_2, \ldots, i_k$ is at most δ, so its Gershgorin circle is within its δ-disk. The Gershgorin radius for any other diagonal element cannot exceed $d + \frac{\delta}{2}$. Therefore the two sets of Gershgorin circles cannot intersect and this proves the theorem.

As an example, let B be the matrix

$$\begin{bmatrix} 2 & 1 & 0 & 1 \\ \epsilon_{21} & 2 & 0 & 0 \\ \epsilon_{31} & \epsilon_{32} & 3 & 1 \\ \epsilon_{41} & \epsilon_{42} & \epsilon_{43} & 3 \end{bmatrix}.$$

Suppose that we wish to bound the error of diagonal elements as eigenvalues by the amount $\delta = .01$. We have $d = .98$, upper $(B) = 2$, $t = 2$, $q = .49$, and $r = .0051$. Then according to Theorem 1, we satisfy this error bound if lower (B) is not greater than $\frac{\delta}{2} r^2 q^3 = 1.53 \cdot 10^{-8}$, which occurs if all ϵ elements are not greater than a third of this value.

When the transformation to near upper triangular form is made using range arithmetic, the real and imaginary parts of diagonal elements are obtained as intervals. Thus a diagonal element will lie in some rectangle in the complex plane, and its δ-disk will lie in a larger rectangle with the same center but with both dimensions increased by 2δ. If several of these augmented rectangles overlap, we can cover them with a still larger rectangle. The following corollary is helpful in dealing with this situation.

COROLLARY 1. *For the n-square matrix B let a number of rectangles $R_1, R_2, \ldots,$ R_s in the complex plane be defined with each rectangle containing the δ-disk for one or* *more diagonal elements of B, with at most t disks to a rectangle. Let d equal the minimum distance between any two of these rectangles. If the inequality in the statement of Theorem 1 is true with this new interpretation of d and t, then every eigenvalue of B lies in a rectangle, and counting multiplicities, every rectangle containing k disks contains k eigenvalues.*

This result follows by repeating the steps of the proof of Theorem 1.

3. Near upper quasi-triangular form by the QR algorithm. This was the first method we tried programming in range arithmetic. By a series of transformations the matrix A is first brought to upper Hessenberg form A' and then to upper quasi-triangular form A'', which allows both 1-square and 2-square blocks along the diagonal [7, pp. 368–379]; the eigenvalues of the diagonal blocks are those of the original matrix A. By permitting 2-square blocks the method avoids complex arithmetic.

Theorem 1 in the preceding section is easily adapted to deal with real matrices in near upper quasi-triangular form, which are produced by a range arithmetic implementation of QR. For this purpose we define two new quantities related to our previous upper (B) and lower (B):

$$\underline{\text{lower}}\,(B) = 3\max_i\left(\sum_{j=1}^{i-1\text{ or }i-2}|b_{ij}|\right),$$

where the sum runs to $i-2$ if $b_{i,i-1}$ is part of a diagonal 2-square block, and otherwise to $i-1$,

$$\underline{\text{upper}}\,(B) = \max_i\left(|\gamma_i| + 3\sum_{i+1\text{ or }i+2}^{n}|b_{ij}|\right),$$

where the sum starts at $i+2$ if $b_{i,i+1}$ is part of a diagonal 2-square block, and otherwise at $i+1$. The term $|\gamma_i|$ is zero unless row i coincides with the upper row of a diagonal 2-square block, and then is given by

$$|\gamma_i| = \begin{cases} |b_{i,i+1} - b_{i+1,i}| & \text{if the block eigenvalues are real,} \\ \sqrt{(b_{i,i} - b_{i+1,i+1})^2 + (b_{i,i+1} + b_{i+1,i})^2} & \text{if the block eigenvalues are complex.} \end{cases}$$

COROLLARY 2. *Assume that the n-square matrix B is real and that an arbitrary but fixed structure of 1-square or 2-square diagonal blocks is associated with B. The n δ-disks are now formed with centers equal to the eigenvalues of these diagonal blocks. The conclusion of Theorem 1 remains true provided we replace the quantities upper(B) and lower(B) by the quantities $\underline{\text{upper}}(B)$ and $\underline{\text{lower}}(B)$, respectively.*

This result follows simply by applying a sequence of 2-square unitary similarity transformations to B in order to triangularize each 2-square diagonal block. The resulting matrix B' is then handled by Theorem 1.

When applying one of these transformations, a typical 2-square diagonal block

$$\begin{bmatrix} b_{i,i} & b_{i,i+1} \\ b_{i+1,i} & b_{i+1,i+1} \end{bmatrix} \quad \text{becomes} \quad \begin{bmatrix} \lambda_1 & \gamma_i \\ 0 & \lambda_2 \end{bmatrix},$$

and the equation for $|\gamma_i|$, already given, may be derived from the relation for unitary transformations

$$|b_{i,i}|^2 + |b_{i,i+1}|^2 + |b_{i+1,1}|^2 + |b_{i+1,i+1}|^2 = |\lambda_1|^2 + |\lambda_2|^2 + |\gamma_i|^2.$$

When the unitary transformations are performed, this may increase the magnitude of some of the matrix elements not in the diagonal blocks. An extreme case may occur when some 2-square block composed of the four elements $b_{q,r}$, $b_{q,r+1}$, $b_{q+1,r}$, $b_{q+1,r+1}$ gets multiplied on the left and on the right by unitary matrices to become the block with elements $b'_{q,r}$, $b'_{q,r+1}$, $b'_{q+1,r}$, $b'_{q+1,r+1}$. We have

$$\begin{aligned}|b'_{q,r}| + |b'_{q,r+1}| &\le \sqrt{2}\cdot\sqrt{|b'_{q,r}|^2 + |b'_{q,r+1}|^2} \quad \text{by the Cauchy inequality}\\ &\le \sqrt{2}\cdot\sqrt{|b'_{q,r}|^2 + |b'_{q+1,r}|^2 + |b'_{q,r+1}|^2 + |b'_{q+1,r+1}|^2}\\ &= \sqrt{2}\cdot\sqrt{|b_{q,r}|^2 + |b_{q+1,r}|^2 + |b_{q,r+1}|^2 + |b_{q+1,r+1}|^2}\\ &\le \sqrt{2}\cdot(|b_{q,r}| + |b_{q+1,r}| + |b_{q,r+1}| + |b_{q+1,r+1}|).\end{aligned}$$

This explains the factor of 3 (which could be replaced by $2\sqrt{2}$) in the definitions of $\underline{\text{upper}}\,(B)$ and $\underline{\text{lower}}\,(B)$.

The implementation of the QR method is somewhat more involved than it would be in floating-point arithmetic. First, we cannot generally guarantee a reduction of A to pure upper Hessenberg form. One has to be content with a near Hessenberg matrix A', which may contain columns of approximate zeros (ranged numbers containing zero in their interval) in its lower triangular part. This is because, during the course of reducing a matrix to upper Hessenberg form, one repeatedly transforms vectors of the form $[*,*,\dots,*]^T$ to the new form $[*,0,\dots,0]^T$, but the transformation is impossible to specify if all elements in the initial vector are approximate zeros, so in this case the vector must be left unchanged.

Second, the question of when to *decouple* [5, p. 235] should be based on a sound estimate of how small $\underline{\text{lower}}\,(A'')$ needs to be to allow printing of the eigenvalues to the desired accuracy, using Corollary 2, which in turn requires some knowledge of the eigenvalues of A (so that one can estimate various parameters such as d and t in Theorem 1). To obtain at least crude approximations of the eigenvalues, the entire problem is first solved in low precision (using an ad-hoc strategy for decoupling) and the results are subsequently used as a guide for when to decouple during the second approach. Third, because the elements below the subdiagonal are not generally zero, the orthogonal similarity transformations need to be applied to the full matrix (or, after decoupling, a full leading principal submatrix) rather than to just the nonzero part as in the case of floating-point arithmetic. Here it is interesting to note that during these transformations the matrix tends to "bleed" into its lower triangular part, so that even if we initially start with a true Hessenberg matrix A', we soon encounter approximate zeros or small nonzero elements below the subdiagonal. (This bleeding can be prevented only at the cost of using transformations with significantly larger ranges.) Finally, the procedure is especially troublesome when applied to a defective matrix: here convergence of subdiagonal elements to zero is only linear and many transformations are necessary to reduce these elements sufficiently. Unfortunately, especially high precision is required in this case because otherwise the growth of the ranges of the matrix elements will make it impossible to get the desired accuracy. This situation usually requires several repetitions of the procedure before sufficiently high precision is reached, compared to mostly two to three runs in the nondefective case (including the initial one to get approximate eigenvalues). Whenever the reduction of A to A'' is complete, Corollary 2 is used to make sure that the number of requested decimal places can indeed be printed; otherwise, the procedure is restarted at a higher precision.

4. Near upper triangular form by Danilevsky's method. Danilevsky's method [4, pp. 251–260] consists of attempting, by a series of transformations, to bring the matrix A to companion matrix form C, shown below:

$$C = \begin{bmatrix} 0 & 0 & \dots & 0 & -c_0 \\ 1 & 0 & \dots & 0 & -c_1 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -c_{n-1} \end{bmatrix}.$$

The characteristic polynomial of C, $\det(C - zI)$, is

$$P(z) = (-1)^n(z^n + c_{n-1}z^{n-1} + \ldots + c_1 z + c_0).$$

If the matrix C could always be found, precise values for the eigenvalues of A could be obtained by getting approximations to the roots of $P(z)$ and then determining their error by making use of the ranged coefficients of $P(z)$ [2, pp. 78–86]. There are difficulties, however, which sometimes occur in obtaining the companion matrix C. By amending the usual Danilevsky procedure to cover the cases of difficulty [2, pp. 124–128], we can generally transform the matrix A to a form C', which either equals the matrix C, or else is composed of companion matrices of intermediate sizes C_1, C_2, ..., C_d, all in diagonal position. In the latter case, however, the elements of C' not belonging to the companion submatrices may not all be zero, but instead some of them may be only "approximate zeros," that is, ranged numbers which contain the zero point in their interval. This precludes multiplying together all the characteristic polynomials of the submatrices C_j to obtain a correctly ranged characteristic polynomial of A, and using this to find the error of root approximations. Instead we use the root approximations only to continue the transformation process toward a near upper triangular form.

We then review how to transform the companion matrix C. If λ is a root of $P(z)$, the row vector

$$X = [1, \lambda, \lambda^2, \ldots, \lambda^{n-1}]$$

is a left eigenvector of C, that is,

$$XC = \lambda X. \tag{1}$$

If λ has multiplicity m, then taking derivatives of (1) with respect to λ leads to the equations

$$X^{(j)}C = \lambda X^{(j)} + jX^{(j-1)}, \qquad j = 1, 2, \ldots, m-1. \tag{2}$$

If we set $X_0 = X$ and $X_j = \frac{1}{j!}X^{(j)}$ for $j = 1, 2, \ldots, m-1$, then (2) shows that the row vector

$$X_j = \left[0, 0, \ldots, 0, \binom{j}{j}, \binom{j+1}{j}\lambda, \ldots, \binom{n-1}{j}\lambda^{n-1-j}\right]$$

satisfies the equation

$$X_j C = \lambda X_j + X_{j-1} \qquad \text{for } j = 1, 2, \ldots, m-1.$$

Therefore, if U is an n-square matrix such that m successive rows are taken equal to $X_{m-1}, X_{m-2}, \ldots, X_0$ for each root λ of multiplicity m, the matrix $J = UCU^{-1}$ gives the Jordan canonical form of C, with each eigenvalue λ of multiplicity m contributing a diagonal m-square block with $m - 1$ ones just above the m diagonal elements λ.

In the general case of the matrix C', we proceed as follows. For each submatrix C_j we obtain all the roots of its characteristic polynomial by the Bairstow method, and their multiplicity by the Euclidean algorithm applied to polynomials [2, pp. 63–67]. The root approximations have their ranges reset to zero and no attempt is made to correctly range them. Their computed multiplicity is not necessarily correct, but in any case, counting computed multiplicities, the number of roots matches the degree of the characteristic polynomial of C_j. There is no difficulty in constructing the matrix U_j and a correctly ranged inverse U_j^{-1} needed to transform C_j to near upper triangular form. After we transform C' to near upper triangular form J' by using all the matrices U_j and their inverses, the diagonal elements can be tested by the method of §2 to determine if the number of correct decimal places is satisfactory, and if not, the whole procedure can be repeated at a higher precision.

The program following this procedure was much superior to the program using the QR procedure, for two reasons. Finding roots by solving a polynomial is clearly more efficient than finding these roots by repeated matrix transformations. Also, by using fewer matrix transformations to achieve the final near upper triangular form, a lower precision of computation is needed to get diagonal elements with an assigned number of mantissa digits and subdiagonal elements that are sufficiently small.

REFERENCES

[1] O. Aberth and M. J. Schaefer, *Precise computation using range arithmetic, via C++*, Trans. Math. Software, to appear.

[2] O. Aberth, *Precise Numerical Analysis*, Wm. C. Brown Publishers, Dubuque, IA, 1988.

[3] G. Alefeld and J. Herzberger, *Introduction to Interval Computation*, Jon Rokne, trans., Academic Press, New York, 1983.

[4] D.K. Faddeeva and V.N. Faddeeva, *Computational Methods of Linear Algebra*, W.H. Freeman and Co., San Francisco, 1963.

[5] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, 1989.

[6] R.E. Moore, *Methods and Applications of Interval Analysis*, SIAM Studies in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1979.

[7] G.W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[8] J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, U.K., 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 242–252, January 1993

019

ON FINDING SUPERNODES FOR SPARSE MATRIX COMPUTATIONS*

JOSEPH W. H. LIU†, ESMOND G. NG‡, AND BARRY W. PEYTON‡

Abstract. A simple characterization of fundamental supernodes is given in terms of the row subtrees of sparse Cholesky factors in the elimination tree. Using this characterization, an efficient algorithm is presented that determines the set of such supernodes in time proportional to the number of nonzeros and equations in the original matrix. Experimental results verify the practical efficiency of this algorithm.

Key words. sparse Cholesky factorizations, elimination trees, supernodes

AMS(MOS) subject classifications. 65F05, 65F50, 65W

1. Introduction.

In the Cholesky factorization of large sparse symmetric positive-definite matrices, factor columns with the same sparsity structure are often clustered together to form a storage or computational unit. Such a grouping of columns has been referred to in the literature as a *supernode.* The term "supernode" first appeared in [4], although the basic idea behind the term was used much earlier.

Indeed, the use of supernodes is found in many algorithms and techniques connected with sparse Cholesky factorization. The related notion of *indistinguishable node subset* is used to accelerate the popular fill-reducing minimum-degree ordering algorithm [9]. The success of the column-compressed factor storage scheme by Sherman [18] can be attributed mainly to the existence of many nontrivial supernodes in most sparse Cholesky factors. Duff and Reid [7] use the term *supervariables* to refer to supernodes, and they explore their use in the multifrontal method. The technique of *static condensation* in finite element applications can also be viewed as the elimination of columns/nodes associated with a supernode.

In [4], Ashcraft et al. discuss supernodal implementations of general sparse and multifrontal factorization methods for vector supercomputers. In their work, supernodes play an important role in achieving a high level of vectorization on such machines. In sparse factorization on shared-memory or distributed-memory parallel machines, the use of supernodes is also central in obtaining good performance. The recent compute-ahead implementation of the fan-in distributed sparse scheme [2] relies quite heavily on the use of supernodes. Preliminary results of work using supernodes in sparse Cholesky factorization on shared-memory parallel machines [14] indicate that substantial improvements in performance can be obtained. The use of clique trees in speeding up a reordering algorithm for parallel elimination [10] is also related to the notion of supernodes. The merits of exploiting supernodes in sparse factorization on modern workstations have been considered by Rothberg and Gupta [16].

The purpose of this paper is to provide an efficient algorithm that determines a set of supernodes of a given sparse matrix using only the structure of the original matrix

* Received by the editors August 25, 1990; accepted for publication (in revised form) May 17, 1991. This research was supported in part by Natural Sciences and Engineering Research Council of Canada grant A5509, and in part by the Applied Mathematical Sciences Research Program, Office of Energy Research, U.S. Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

† Department of Computer Science, York University, North York, Ontario, Canada M3J 1P3 (joseph@cs.yorku.ca).

‡ Mathematical Sciences Section, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6367 (esmond@msr.epm.ornl.gov and peyton@msr.epm.ornl.gov).

and its elimination tree. Given the many uses of supernodes, this scheme should be a useful addition to practical sparse matrix software packages. The algorithm allows a more modularized, and hence flexible, approach to the symbolic processing phase of the solution process. This flexibility should become more valuable as developers of production-level software contend with a growing array of complex computer architectures.

An outline of this paper follows. In §2, we formally define the notions of supernodes and so-called "fundamental" supernodes [3]. We then use the column and row structures of the sparse Cholesky factor matrix to characterize these supernodes. An efficient algorithm for finding fundamental supernodes is described in §3. We show, moreover, that the execution time of the algorithm is linear with respect to the number of nonzeros and equations in the original given matrix. Section 4 contains timing results that demonstrate the practical efficiency of the algorithm. Section 5 contains our concluding remarks and discusses some possible applications of the algorithm.

2. Supernodes and fundamental supernodes.

2.1. Cholesky factor characterization of supernodes. We assume that the reader is familiar with the terminology used in the study of sparse elimination, such as fill, ordering, and related concepts; moreover, we assume familiarity with the use of graphs to model the sparsity structure of matrices. All the necessary material can be found in [5] or [8]. We also assume familiarity with the definitions and relevance of the *elimination tree* structure and tree *postordering* in sparse elimination. Details on elimination trees and postorderings can be found in [13].

Let A be a given $n \times n$ sparse symmetric positive-definite matrix, with L as its Cholesky factor. The (i,j)th entries of A and L are represented by a_{ij} and ℓ_{ij}, respectively. Without loss of generality, the matrix A is assumed to be irreducible, so that its elimination forest is indeed a tree. Furthermore, we assume that the matrix A has been preprocessed so that its ordering is already a postordering of the elimination tree. (Reasons for this assumption are given later in this section and also in subsequent sections of the paper.) For our discussion, we use T to represent the elimination tree of A, and $T[i]$ to represent the subtree consisting of node i and all its descendants. The nodes in each subtree $T[i]$ are numbered consecutively by the postordering. We let $G(A)$ denote the graph of A, G^* the filled graph associated with L, and $\mathrm{m\,adj}_G(i)$ the set of nodes $j > i$ adjacent to i in G.

We define a *maximal supernode* with respect to the postordering to be a maximal block of contiguous columns of L, whose diagonal block is full lower triangular, and whose off-block-diagonal column sparsity structures are identical. Figure 1 shows the maximal supernodes in the Cholesky factor of a matrix defined by a 7×7 finite element grid ordered by nested dissection; columns between two vertical lines belong to the same maximal supernode. For brevity, we shall use the term "supernode" to refer to a maximal supernode. To formalize this notion, we need the following notation and terminology. For a sparse n-vector v, we define the *structure* of v to be

$$\mathrm{Struct}(v) = \{j \mid v_j \neq 0\}.$$

Furthermore, we use the notation $\eta(v)$ to represent $|\mathrm{Struct}(v)|$, that is, the number of nonzeros in the vector v. Finally, we use "column subset" to refer to a subset of the column subscripts.

Using this terminology, we can express the notion of supernodes as follows. The column subset $S = \{s, s+1, \ldots, s+t-1\}$ is a supernode of the matrix L if and only

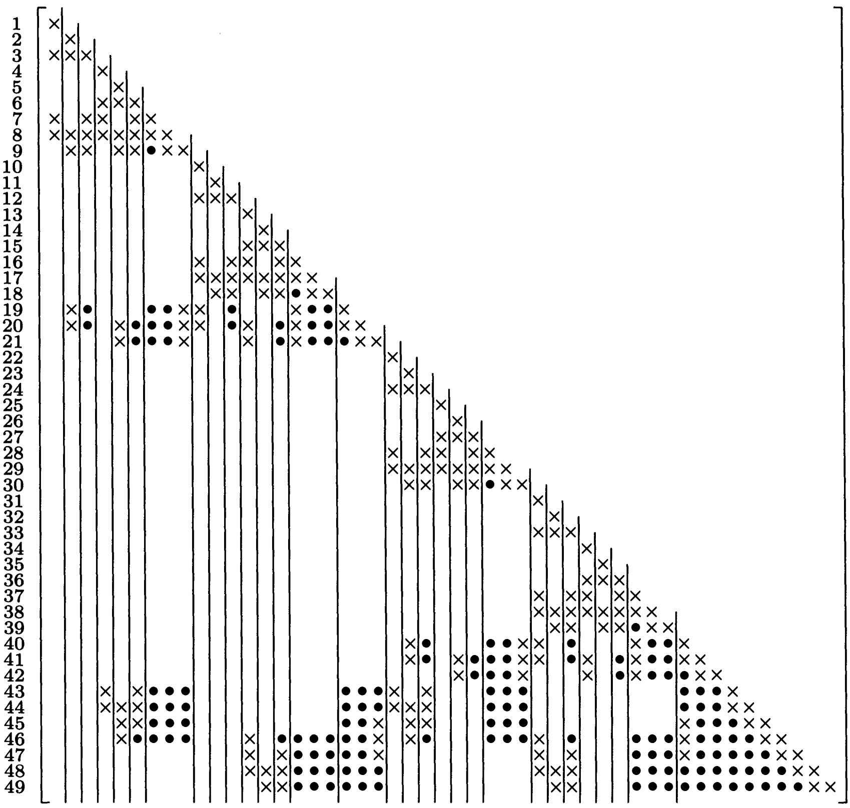


FIG. 1. *Supernodes for the 7 × 7 grid problem ordered by nested dissection.* (× *and* • *refer to nonzeros in A and fill in L, respectively.*)

if it is a maximal contiguous column subset satisfying

$$\operatorname{Struct}(L_{*,s}) = \operatorname{Struct}(L_{*,s+t-1}) \cup \{s, \dots, s+t-2\}.$$

In graph terminology, a supernode is a maximal clique of consecutively numbered nodes $S = \{s, s+1, \dots, s+t-1\}$ for which

$$\operatorname{m\,adj}_{G^*}(s) = \operatorname{m\,adj}_{G^*}(s+t-1) \cup \{s, \dots, s+t-2\}.$$

The maximal property of a supernode implies that neither $S \cup \{s-1\}$ nor $S \cup \{s+t\}$ satisfies this structural condition. We can verify that the set of supernodes shown in Fig. 1 satisfies the maximal property.

The particular set of supernodes obtained by applying this definition depends on the particular postordering used to number the nodes. Indeed, one postordering may give rise to fewer supernodes than another. It is trivial to generalize the definition so that it applies to any topological ordering of the elimination tree,[1] but many of these orderings would define far too many supernodes and therefore capture little of the supernode structure actually available in the factor. In contrast, postorderings define

[1] A topological ordering of an elimination tree is an ordering in which each parent is labelled after its children [13].

supernode sets that in practice take advantage of almost all the redundant column structure to be found in the factor.

Finding the set of supernodes of L is an almost trivial task once the structure of the Cholesky factor L is determined. Indeed, we need only the factor column nonzero counts, rather than factor column structures, to find the set of supernodes. An algorithm using nonzero counts and the elimination tree structure can be formulated based on the following result, which follows easily from the definition of supernodes and the fact that

$$\mathrm{m\,adj}_{G^*}(i) - \{k \mid k \leq j\} \subseteq \mathrm{m\,adj}_{G^*}(j)$$

for any two nodes i and j for which $j \in \mathrm{m\,adj}_{G^*}(i)$. Formal proof of a very similar result appears in [15].

THEOREM 2.1. *The column subset $S = \{s, s+1, \ldots, s+t-1\}$ is a supernode of the matrix L if and only if S is a maximal set of contiguous columns such that $s+i-1$ is a child of $s+i$ in the elimination tree for $i = 1, \ldots, t-1$, and*

$$\eta(L_{*,s}) = \eta(L_{*,s+t-1}) + t - 1.$$

Algorithms that use results much like Theorem 2.1 to find supernodes are presented in [3] and [15]. Our goal is an efficient algorithm that generates the supernodes *before* the symbolic factorization, that is, before the structure of the Cholesky factor is computed. We might try to achieve this goal by applying Theorem 2.1, as in the following scheme.

1. Compute the elimination tree, postorder the tree, and compute the factor column nonzero counts.
2. Use Theorem 2.1 to generate the supernodes.

An algorithm for determining the factor column nonzero counts appears in [13]. The primary problem with this approach is the expense of computing the factor column nonzero counts prior to the symbolic factorization. Indeed, computing the nonzero counts requires $O(n + \eta(L))$ operations. We have implemented this approach, and our tests indicate that this method is much less efficient than the new algorithm introduced in the next two sections. The new approach finds the supernodes directly from A by determining the leaves of the row subtrees, which requires $O(n + \eta(A))$ operations, and has possible application in situations where a more modular approach to the symbolic factorization is desirable.

2.2. Row-subtree characterization of fundamental supernodes. In [3], Ashcraft and Grimes introduce the notion of fundamental supernodes. A *fundamental supernode* is a maximal contiguous column subset $\{s, s+1, \ldots, s+t-1\}$ such that $s+i-1$ is the *only* child of $s+i$ in the elimination tree for $i = 1, \ldots, t-1$ and

$$\mathrm{Struct}(L_{*,s}) = \mathrm{Struct}(L_{*,s+t-1}) \cup \{s, \ldots, s+t-2\}.$$

This restriction on the columns that can appear together in the same supernode is appropriate for several reasons. First, unlike the set of maximal supernodes, the set of fundamental supernodes is necessarily the same regardless of the particular postordering in use. Second, while our new algorithm could be designed to generate maximal rather than fundamental supernodes, the difference between the two is rarely of any practical consequence—typically, very few maximal supernodes are replaced by more than one supernode in the corresponding set of fundamental supernodes. For example, for the matrix shown in Fig. 1, the set of fundamental supernodes

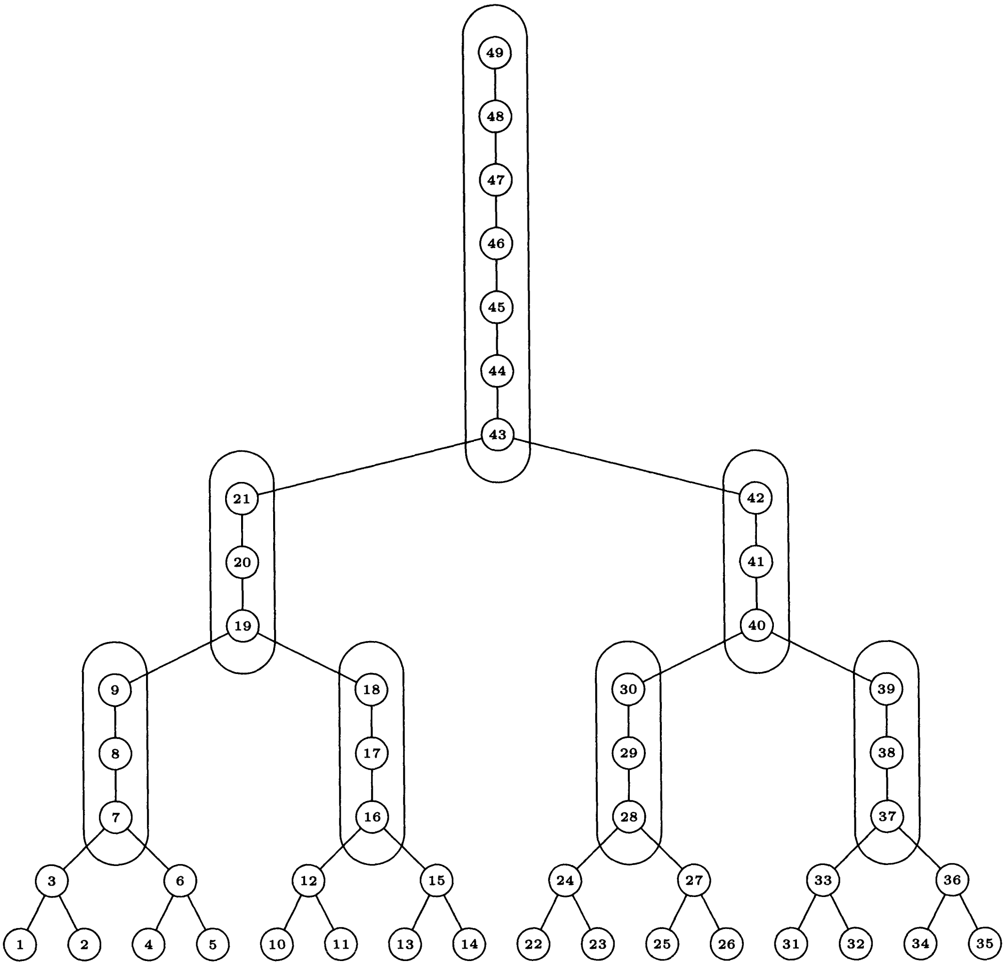


FIG. 2. *Supernodal elimination tree induced by the fundamental supernodes for the matrix shown in Fig. 1. Supernodes containing more than one node are enclosed by ovals; all nodes in the elimination tree not enclosed by ovals are singleton supernodes.*

is the same as the set of maximal supernodes with one exception: the supernode $\{40, 41, 42, 43, 44, 45, 46, 47, 48, 49\}$ is partitioned into two supernodes $\{40, 41, 42\}$ and $\{43, 44, 45, 46, 47, 48, 49\}$ because node 43 has two children in the elimination tree (see Fig. 2). Third, the supernodal version of the elimination tree induced by a set of fundamental supernodes (see Fig. 2) preserves in a strict sense the data dependency results that hold for the nodal elimination tree: that is, for a set of fundamental supernodes, no column in a supernode can be completed until after the completion of all columns belonging to descendant supernodes in the supernodal elimination tree. For a given set of maximal supernodes, it may be the case that some columns in a supernode can be completed before all columns belonging to descendant supernodes are completed. The strict form of data dependency that holds for fundamental supernodes is useful in designing parallel factorization algorithms that make explicit use of supernodes.

Finding the set of maximal or fundamental supernodes is equivalent to finding the first columns of the supernode subsets. In this subsection, we shall provide a characterization of such first columns/nodes of fundamental supernodes in terms of the elimination tree. The characterization uses the row structure of the Cholesky factor.

As before, let T be the elimination tree of the matrix A, and $T[i]$ the subtree rooted at node i. We also use $T[i]$ to represent the node subset associated with the subtree, namely, node i together with all its proper descendants in the elimination tree. The notation $|T[i]|$ is used to denote the number of nodes in the subtree $T[i]$. Therefore, the number of proper descendants of node i is given by $|T[i]| - 1$.

In [17], Schreiber shows that the structure of the ith row of the Cholesky factor L is a *pruned subtree* of the elimination tree rooted at node i. Liu, in [12], gives a complete characterization of the structure of factor row L_{i*}, and calls the corresponding pruned subtree the ith *row subtree*. We quote the following result without proof.

THEOREM 2.2 (see Liu [12]). *Node $j \in T[i]$ is a leaf node of the ith row subtree of L if and only if $a_{ij} \neq 0$ and $a_{ik} = 0$ for every $k \in T[j] - \{j\}$.*

Note that the characterization in Theorem 2.2 is in terms of nonzeros in the original matrix A. We are now ready to characterize fundamental supernodes using row structures and row subtrees of the factor matrix.

THEOREM 2.3. *Column j is the first node in a fundamental supernode if and only if node j has two or more children in the elimination tree, or j is a leaf of some row subtree of the elimination tree of A.*

Proof. "*if*": In the first case, if node j has two or more children in the elimination tree, it is clear from the definitions that it must be the first node of the fundamental supernode to which it belongs. In the second case, assume that node j is a leaf of some row subtree, say the ith row subtree. If j has no children, it must start a fundamental supernode. Therefore, we need consider only the case when j has exactly one child. Its only child must be $j-1$ because of the postordering assumption. Since j is a leaf of the row subtree rooted at i, it follows from Theorem 2.2 that $a_{ij} \neq 0$ and $a_{i,j-1} = 0$, which implies that ℓ_{ij} is structurally nonzero while $\ell_{i,j-1}$ is structurally zero. It then follows that $\text{Struct}(L_{*,j-1}) \subset \text{Struct}(L_{*,j}) \cup \{j-1\}$, and thus node j cannot belong to the same supernode as $j-1$. Node j must, therefore, start a new fundamental supernode.

"*only if*": Assume that node j is the first node of its fundamental supernode. We need to show that if j has one or no children, it is a leaf of some row subtree. The result is obvious if j has no children. Consider the case when j has exactly one child $j-1$. Since j starts a fundamental supernode and has only one child, the maximal condition implies that there exists $i > j$ such that $\ell_{i,j-1} = 0$ and $\ell_{ij} \neq 0$. In other words, the node j belongs to the ith row subtree of L, whereas its only child $j-1$ does not. This is possible only if j is a leaf of this row subtree. □

Theorem 2.3 provides a simple condition to determine the starting node of fundamental supernodes. Computing the number of children for each node in the elimination tree is a trivial task. It can be done in $O(n)$ time by traversing the elimination tree in a postorder sequence. All we need is an efficient algorithm to identify those nodes that are leaves of at least one row subtree in the elimination tree. A simple marking scheme can be used to detect the leaves of successive row subtrees in the algorithm that generates the elimination tree. Implementing this scheme, however, requires removal of the path compression technique [13], which is essential for efficient computation of the elimination tree. We have implemented this scheme, but found the resulting $O(\eta(L))$ algorithm to be too inefficient.

3. A linear-time algorithm for finding fundamental supernodes. Theorem 2.3 gives a simple characterization of fundamental supernodes. To use this characterization, we need to identify the nodes that are leaves of at least one row subtree. In this section, we consider an efficient linear-time algorithm to determine the Boolean vector ISLEAF($*$), where ISLEAF(j) is true if and only if node j is a leaf of some row subtree of the elimination tree T.

The result in Theorem 2.2 can be used to find the leaves of each row subtree. However, it does not immediately lead to a practical algorithm, since determining if a_{ik} is zero for every proper descendant k of j in the subtree $T[j]$ is far too expensive. By making use of the postordering property, Liu [12] gives a more usable form of the same result, which we quote here.

THEOREM 3.1 (see Liu [12]). *Let $c_1 < c_2 < \cdots < c_s < i$ be the column subscripts of all the nonzeros in A_{i*} less than i, that is, the structure of row i of A to the left of the diagonal. Node c_t is a leaf of the ith row subtree if and only if $t = 1$ or $c_{t-1} \notin T[c_t]$.*

The testing of $c_{t-1} \notin T[c_t]$ in Theorem 3.1 can be further simplified by using a property that holds for ancestor-descendant pairs when the nodes are postordered [1].

COROLLARY 3.2 (see Liu [12]). *Let $c_0 < c_1 < \cdots < c_s < i$, where $c_0 = 0$, and $c_1, \ldots, c_s$ are as in Theorem 3.1. Node c_t is a leaf of the ith row subtree if and only if $c_{t-1} < c_t - |T[c_t]| + 1$.*

One way to compute ISLEAF($*$) is to apply Corollary 3.2 to the rows of A one by one, determining the set of leaves for each row subtree and flagging the nodes accordingly. However, this involves *sorting* the column subscripts of nonzeros in A_{i*} to the left of its diagonal entry for every row of A. Although this can be accomplished with a bin sort in time proportional to $|A|$, our experience indicates that it is necessary to avoid such sorting in order to obtain a method whose practical efficiency is acceptable.

The following observation can be used to avoid such sorting: j is a leaf of some row subtree of the elimination tree of A if and only if for some $i \in \mathrm{madj}_{G(A)}(j)$, we have $i \notin \mathrm{madj}_{G(A)}(k)$ for all $k \in T[j] - \{j\}$. Using this observation and the postordering, we obtain an algorithm that uses a simple new marking scheme to avoid the undesirable sorting (see Fig. 3). In this algorithm, the columns of the matrix A are processed in postorder. While processing column A_{*j}, for each nonzero a_{ij} below the diagonal, the algorithm checks to see if node j is a leaf of the ith row subtree (using Corollary 3.2). A temporary integer vector PREV_ROWNZ(*) is used to remember the location of the latest nonzero encountered for each row. It is also necessary to compute the subtree sizes $|T[j]|$, which can be done in $O(n)$ time. It is obvious that the overall algorithm runs in $O(n+\eta(A))$ time. However, the linear-time complexity of the algorithm is obtained by exploiting a postordering of the elimination tree, which can be obtained only by first computing the elimination tree itself. It is worth noting that the algorithm for this task has greater complexity than the new algorithm; it is *almost linear* in $\eta(A)$ [13].

4. Timing results. In this section, we provide results of some experiments verifying the practical efficiency of the algorithm in Fig. 3 for computing fundamental supernodes. The experiments were performed on a Sun 3/80 workstation with eight Mbytes of main memory. The programs were written in Fortran and compiled using the Sun Fortran compiler with optimization turned on. The test problems used in the experiments were selected from the Harwell–Boeing Sparse Matrix Collection [6], and they are listed in Table 1. Structural and timing statistics are provided in Tables 2

```
for column $j$ := 1 to $n$ do
    ISLEAF($j$) := false;
    PREV_ROWNZ($j$) := 0 ;
end for

for node $j$ := 1 to $n$ do
    compute $|T[j]|$ ;
end for

for column $j$ := 1 to $n$ do
    for each $a_{ij} \neq 0$ with $i > j$ do
        $k$ := PREV_ROWNZ($i$) ;
        if $k < j - |T[j]| + 1$ then
            ISLEAF($j$) := true;
        endif;
        PREV_ROWNZ($i$) := $j$ ;
    end for
end for
```

FIG. 3. *Determination of row-subtree leaves in the elimination tree.*

and 3, respectively. Each matrix was reordered by the minimum-degree algorithm. Three times are reported for each matrix: the time to compute the elimination tree and its postordering, the time to compute the supernodes with the new algorithm, and the time to compute the symbolic factorization using the routine `smbfct` from SPARSPAK.

TABLE 1
List of test problems.

Problem	Description
BCSPWR10	Structure of representation of entire U.S. power network
BCSSTK13	Stiffness matrix, fluid flow generalized eigenvalues
BCSSTK14	Stiffness matrix—roof of Omni Coliseum, Atlanta
BCSSTK15	Stiffness matrix—module of an offshore platform
BCSSTK16	Stiffness matrix—Corps of Engineers dam
BCSSTK17	Stiffness matrix—elevated pressure vessel
BCSSTK18	Stiffness matrix—R.E.Ginna nuclear power station
BCSSTK21	Stiffness matrix—clamped square plate
BCSSTK23	Stiffness matrix—portion of a three-dimensional globally triangular bldg.
BCSSTK24	Stiffness matrix—winter sports arena
BCSSTK25	Stiffness matrix—76-story skyscraper
BCSSTK28	Solid element model (MSC NASTRAN)
CEGB2802	Finite element problem. Turbine blade
CEGB3306	Finite element framework problem, essentially in two dimensions
CEGB2919	Finite element problem. Three-dimensional cylinder with flange
CEGB3024	Finite element problem. Two-dimensional reactor core section
LOCK2232	Finite element problem. Lockheed tower problem
LOCK3491	Finite element problem. Lockheed cross-cone problem

The timing statistics in Table 3 demonstrate the practical efficiency of the new

TABLE 2
Structural statistics on test problems.

Problem	n	$\eta(A)$	$\eta(L)$	Number of supernodes
BCSPWR10	5,300	21,842	28,064	4,960
BCSSTK13	2,003	83,883	271,671	599
BCSSTK14	1,806	63,454	112,267	503
BCSSTK15	3,948	117,816	651,222	1,295
BCSSTK16	4,884	290,378	741,178	691
BCSSTK17	10,974	428,650	1,005,859	2,595
BCSSTK18	11,948	149,090	662,725	7,438
BCSSTK21	3,600	26,600	90,454	2,420
BCSSTK23	3,134	45,178	420,311	1,522
BCSSTK24	3,562	159,910	278,922	414
BCSSTK25	15,439	252,241	1,416,568	7,288
BCSSTK28	4,410	219,024	346,894	454
CEGB2802	2,802	277,470	267,972	271
CEGB3306	3,306	75,000	68,205	599
CEGB2919	2,919	321,603	375,642	252
CEGB3024	3,024	79,876	115,702	685
LOCK2232	2,232	80,376	73,308	318
LOCK3491	3,491	160,519	236,339	479

TABLE 3
Timing statistics (in seconds).

Problem	Tree and postord.	Supernodes	Symbolic factorization (smbfct)
BCSPWR10	0.20	0.16	0.58
BCSSTK13	0.36	0.22	1.10
BCSSTK14	0.28	0.18	0.68
BCSSTK15	0.56	0.32	1.46
BCSSTK16	1.26	0.74	3.74
BCSSTK17	1.90	1.14	5.06
BCSSTK18	0.90	0.56	2.22
BCSSTK21	0.20	0.12	0.56
BCSSTK23	0.26	0.16	0.76
BCSSTK24	0.68	0.42	1.74
BCSSTK25	1.44	0.82	3.58
BCSSTK28	0.92	0.56	2.40
CEGB2802	1.12	0.68	4.62
CEGB3306	0.34	0.22	0.72
CEGB2919	1.28	0.78	5.34
CEGB3024	0.38	0.22	0.80
LOCK2232	0.34	0.22	0.74
LOCK3491	0.68	0.42	1.74

algorithm for finding the fundamental supernodes. The elimination tree and a postordering must be computed before the fundamental supernodes can be computed with the new algorithm. In every case, the time spent computing the supernodes is less than the time spent computing the elimination tree and postordering, as might be expected. Also note that the time to compute the supernodes is a very modest fraction of the time to compute the symbolic factorization using the routine `smbfct` from SPARSPAK. Based on these timings, it appears that the new algorithm for finding the fundamental supernodes is potentially useful in cases where the supernodes are needed *before* the symbolic factorization is completed.

5. Concluding remarks. We have introduced a new characterization of fundamental supernodes that depends on the structure of A and its elimination tree, and requires no knowledge of the structure of L. We have used this characterization to devise an efficient algorithm for determining the fundamental supernodes of a sparse symmetric matrix *prior to* the symbolic factorization step.

A possible application of this scheme is the development of the following two-step method to perform the symbolic factorization.

1. Compute the supernodes with the new algorithm.
2. Compute a *supernodal symbolic factorization* that takes advantage of access to the supernodes to improve its efficiency.

We have implemented this two-step method to perform the symbolic factorization. Preliminary experiments were reported in [11], and the results demonstrated that the two-step approach was much more efficient than performing the symbolic factorization with the routine `smbfct` from SPARSPAK. However, it was pointed out by Dr. Stan Eisenstat of Yale University and an anomymous referee that the performance of `smbfct` could be improved by reorganizing the computation to eliminate some unnecessary sorting operations, and by removing a type of subscript compression not computed by the supernodal symbolic factorization. Additional experiments indicate that the improved version of `smbfct` was somewhat more efficient than the two-step method, though the two approaches are quite comparable in most cases. Thus, it appears that the two-step approach may not be appropriate for *general* symbolic factorization.

Nonetheless, the more modular approach has desirable features and may be preferable in certain situations, such as the following:

- Parallel symbolic factorization on distributed-memory parallel machines.
- Implementation of the supernode amalgamating strategy introduced in [3], which has been used in three multifrontal codes that we are aware of: Ashcraft and Grimes (Boeing), Chao and Vu (Cray Research Inc.), and Amestoy and Duff (CERFACS).
- Production quality sparse matrix software, where a tight bound on the space required for the symbolic factorization may be needed before that computation can begin.
- Production quality sparse matrix software, where some ordering modules generate the elimination tree and the supernodes as a natural by-product (e.g., minimum degree), while others currently do not (e.g., automatic nested dissection).

It is worth noting that even though the two-step symbolic factorization appears to be slightly less efficient than its one-step counterpart, each of the two routines used in the two-step approach is much simpler than `smbfct` and its variants, which are fairly intricate codes.

Acknowledgments. The authors would like to thank Dr. Stan Eisenstat and one of the referees for pointing out how to reorganize the symbolic factorization routine `smbfct` from SPARSPAK in order to obtain substantial improvements in its performance. They also thank the referees for many thoughtful suggestions, which improved the presentation of the material.

REFERENCES

[1] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, Reading, MA, 1983.

[2] C. Ashcraft, S. Eisenstat, J. Liu, B. Peyton, and A. Sherman, *A compute-ahead implementation of the fan-in sparse distributed factorization scheme*, Tech. Rep. CS-90-03, Dept. of Computer Science, York Univ., Ontario, Canada, 1990.

[3] C. Ashcraft and R. Grimes, *The influence of relaxed supernode partitions on the multifrontal method*, ACM Trans. Math. Software, 15 (1989), pp. 291–309.

[4] C. Ashcraft, R. Grimes, J. Lewis, B. Peyton, and H. Simon, *Progress in sparse matrix methods for large linear systems on vector supercomputers*, Internat. J. Supercomput. Appl., 1 (1987), pp. 10–29.

[5] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*, Oxford University Press, Oxford, U.K., 1987.

[6] I. S. Duff, R. Grimes, and J. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[7] I. S. Duff and J. K. Reid, *The multifrontal solution of indefinite sparse symmetric linear equations*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.

[8] J. A. George and J. W. H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1981.

[9] ———, *The evolution of the minimum degree ordering algorithm*, SIAM Rev., 31 (1989), pp. 1–19.

[10] J. G. Lewis, B. W. Peyton, and A. Pothen, *A fast algorithm for reordering sparse matrices for parallel factorization*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1146–1173.

[11] J. Liu, E. Ng, and B. Peyton, *On finding supernodes for sparse matrix computations*, Tech. Rep. ORNL/TM-11563, Mathematical Sciences Section, Oak Ridge National Laboratory, Oak Ridge, TN, 1990.

[12] J. W. H. Liu, *A compact row storage scheme for Cholesky factors using elimination trees*, ACM Trans. Math. Software, 12 (1986), pp. 127–148.

[13] ———, *The role of elimination trees in sparse factorization*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 134–172.

[14] E. Ng and B. Peyton, *A supernodal Cholesky factorization algorithm for shared-memory multiprocessors*, Tech. Rep. ORNL/TM-11814, Mathematical Sciences Section, Oak Ridge National Laboratory, Oak Ridge, TN, 1991.

[15] A. Pothen, *Simplicial cliques, shortest elimination trees, and supernodes in sparse Cholesky factorization*, Tech. Rep. CS-88-13, Dept. of Computer Science, The Pennsylvania State Univ., University Park, PA, 1988.

[16] E. Rothberg and A. Gupta, *Efficient sparse matrix factorization on high-performance workstations—exploiting the memory hierarchy*, ACM Trans. Math. Software, 17 (1991), pp. 313–334.

[17] R. Schreiber, *A new implementation of sparse Gaussian elimination*, ACM Trans. Math. Software, 8 (1982), pp. 256–276.

[18] A. H. Sherman, *On the efficient solution of sparse systems of linear and nonlinear equations*, Ph.D. thesis, Dept. of Computer Science, Yale Univ., New Haven, CT, 1975.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 253–258, January 1993

020

A NOTE ON NESTED DISSECTION FOR RECTANGULAR GRIDS*

M. V. BHAT†, W. G. HABASHI‡, J. W. H. LIU§, V. N. NGUYEN†, AND M. F. PEETERS¶

Abstract. A new ordering scheme is presented for sparse matrices associated with rectangular grid problems. The new scheme combines features of both the nested dissection and natural orderings to obtain orderings superior to either of these methods for such problems. Asymptotic bounds and experimental results on sparse factorization operations are provided to demonstrate the advantages of the new ordering.

Key words. sparse matrix, reordering, nested dissection, grid

AMS(MOS) subject classifications. 65F50, 65F25

1. Introduction.

It is well known that matrix ordering can have a dramatic impact on the storage and computational costs in the direct solution of sparse symmetric positive definite linear systems. For sparse matrices associated with regular square grid problems, the nested dissection ordering by George [3] is known to give asymptotically optimal results in both storage and operation counts in factorization.

In this paper, we describe a new ordering appropriate for rectangular grid problems. The scheme orders a set of rectangular subgrids *locally* by the standard nested dissection ordering. The remaining nodes are numbered in a manner similar to the natural ordering of grids. It can be viewed as combining the advantages of the nested dissection and natural orderings. Asymptotic bounds and experimental results are used to demonstrate that significant reductions in arithmetic operations can be achieved. It is also observed that this ordering can provide insight into finding better orderings for general sparse matrix problems, and it can be a useful example to study when considering parallel sparse matrix factorization schemes.

2. Motivation: Natural and nested dissection orderings for rectangular grids.

Consider a given rectangular grid with a nine-point operator. Let the grid be h-by-k. Without loss of generality, we assume that $h \geq k$. The *natural* ordering is given by numbering the nodes on successive horizontal grid lines (the narrow side), as shown in Fig. 1. This ordering is well suited for band-type sparse schemes.

On the other hand, the well-known *nested dissection* ordering by George [1], [3], [4] is more suited for sparse schemes that take advantage of all possible nonzeros. This scheme is known to be asymptotically optimal in terms of fills and factorization operations when the grid is square. The standard scheme can be adapted for rectangular grids by choosing the dissecting grid line on the *narrower* side of the grid. Therefore, the first dissector for the given h-by-k grid is the middle horizontal grid line. The overall ordering is obtained by recursively dissecting each subgrid on the narrower side of the rectangle, and numbering nodes of the dissector last in the ordering. In

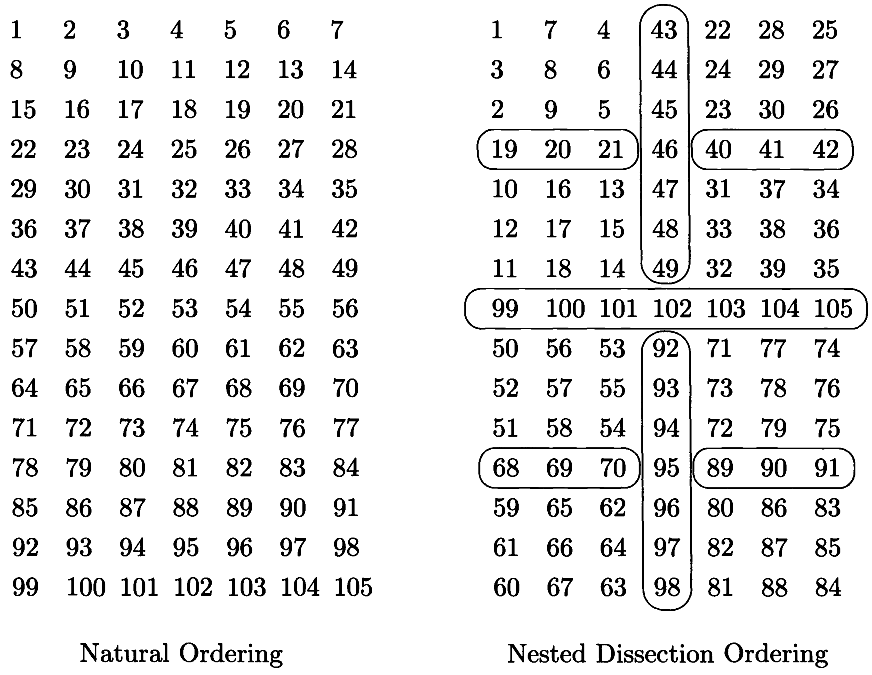
* Received by the editors March 1, 1990; accepted for publication (in revised form) May 17, 1991.

† Pratt and Whitney Canada, Longueuil, Quebec, Canada J4G 1A1.

‡ Department of Mechanical Engineering, Concordia University, Montreal, Quebec, Canada H3G 1M8. This author's research was supported in part by Natural Sciences and Engineering Research Council of Canada operating grant OGPIN 013 and strategic grant STREQ 040.

§ Department of Computer Science, York University, North York, Ontario, Canada M3J 1P3 (joseph@cs.yorku.ca). This author's research was supported in part by Natural Sciences and Engineering Research Council of Canada grant A5509.

¶ Pratt and Whitney Canada, Mississauga, Ontario, Canada, L5T 1J3.



1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	32	33	34	35
36	37	38	39	40	41	42
43	44	45	46	47	48	49
50	51	52	53	54	55	56
57	58	59	60	61	62	63
64	65	66	67	68	69	70
71	72	73	74	75	76	77
78	79	80	81	82	83	84
85	86	87	88	89	90	91
92	93	94	95	96	97	98
99	100	101	102	103	104	105

Natural Ordering

1	7	4	43	22	28	25
3	8	6	44	24	29	27
2	9	5	45	23	30	26
19	20	21	46	40	41	42
10	16	13	47	31	37	34
12	17	15	48	33	38	36
11	18	14	49	32	39	35
99	100	101	102	103	104	105
50	56	53	92	71	77	74
52	57	55	93	73	78	76
51	58	54	94	72	79	75
68	69	70	95	89	90	91
59	65	62	96	80	86	83
61	66	64	97	82	87	85
60	67	63	98	81	88	84

Nested Dissection Ordering

FIG. 1. *Natural and nested dissection orderings for* 15*-by-*7 *grid.*

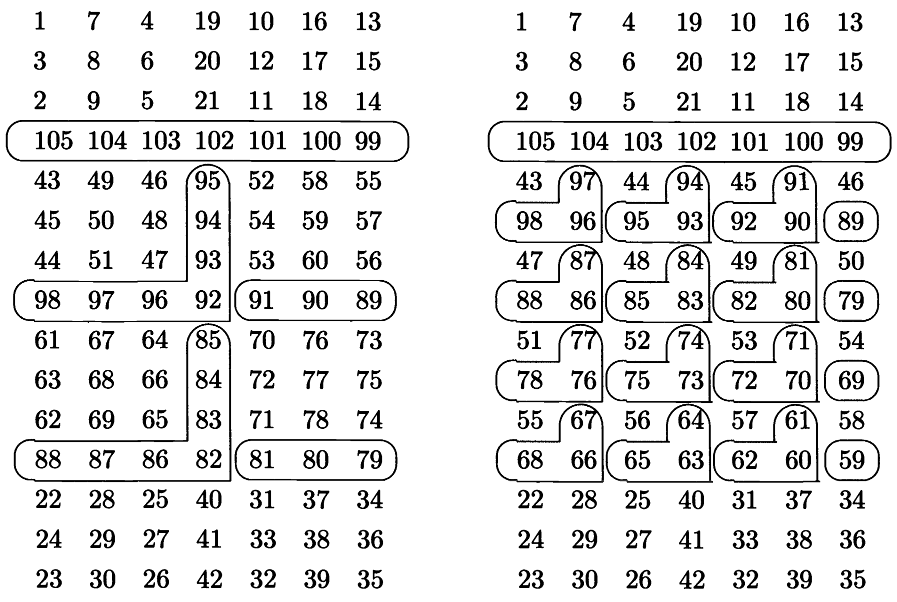
Fig. 1, we illustrate the natural and nested dissection orderings for the 15-by-7 grid problem.

Table 1 contains statistics on the natural and nested dissection orderings for a sequence of h-by-15 grids. The columns labeled "Nonz." give the number of off-diagonal nonzeros of the corresponding sparse Cholesky factor, and "Opn. Count" the number of multiplicative operations to compute the factor. It is interesting to note that the natural ordering is better than the nested dissection ordering in operations for $h \geq 50$, and subsequently better in the number of factor zeros for $h \geq 100$. Note that for $h = 300$, there is a saving of more than 30 percent in operations if the natural ordering is used.

TABLE 1
*Factorization statistics for h-by-*15 *rectangular grids.*

h	Natural ordering		Nested dissection	
	Nonz.	Opn. count	Nonz	Opn. count
15	3360	31164	2553	21327
50	11725	110369	11003	121056
100	23675	223519	23759	283918
300	71475	676119	76156	974396

The experimental results in Table 1 can be explained easily from analysis results in the literature. We only address the differences in factorization operation counts. For the natural ordering on a given h-by-k grid, it is easy to verify that the sparse Cholesky factorization of such matrices requires $hk^3/2+O(hk^2)$ arithmetic operations. For the standard nested dissection ordering (where the narrower side is dissected), an asymptotic bound of $18.99hk^2 + O(k^3)$ on the arithmetic operations can be obtained.



1	7	4	19	10	16	13
3	8	6	20	12	17	15
2	9	5	21	11	18	14
105	104	103	102	101	100	99
43	49	46	95	52	58	55
45	50	48	94	54	59	57
44	51	47	93	53	60	56
98	97	96	92	91	90	89
61	67	64	85	70	76	73
63	68	66	84	72	77	75
62	69	65	83	71	78	74
88	87	86	82	81	80	79
22	28	25	40	31	37	34
24	29	27	41	33	38	36
23	30	26	42	32	39	35

With Partition Factor p=2

1	7	4	19	10	16	13
3	8	6	20	12	17	15
2	9	5	21	11	18	14
105	104	103	102	101	100	99
43	97	44	94	45	91	46
98	96	95	93	92	90	89
47	87	48	84	49	81	50
88	86	85	83	82	80	79
51	77	52	74	53	71	54
78	76	75	73	72	70	69
55	67	56	64	57	61	58
68	66	65	63	62	60	59
22	28	25	40	31	37	34
24	29	27	41	33	38	36
23	30	26	42	32	39	35

With Partition Factor p=4

FIG. 2. *Local nested dissection orderings for* 15*-by-*7 *grid.*

Using these bounds, we note that for any value of $k < 37$ and relatively large h, the nested dissection ordering requires more factorization operations than the natural ordering. This explains the results in Table 1.

3. A local nested dissection ordering for rectangular grids. The observations in the last section provide motivation to consider a new ordering scheme, which combines the advantages of both the natural and nested dissection orderings for grid problems. As before, let the given grid be h-by-k, where $h \geq k$. The new scheme orders the two rectangular subgrids consisting of the first $k/2$ rows and the last $k/2$ rows, respectively, of the grid using the standard nested dissection ordering. The remaining $h-k$ rows of the grid are numbered as follows. Consider a partitioning factor p, with $1 \leq p < k$. Partition the remaining $(h-k)$-by-k rectangular grid into smaller square subgrids of size approximately k/p, by a set of horizontal and vertical grid lines. Horizontally, there are p partitions; vertically, there are $p(h-k)/k$ partitions. This gives $p^2(h-k)/k$ square subgrids of size k/p.

The new ordering numbers the nodes associated with each square subgrid by the standard nested dissection ordering. The remaining nodes associated with the partitioning grid lines are then numbered using a scheme similar to the natural ordering. The only difference is that a set of nodes associated with the boundary of a square subgrid is numbered consecutively. To illustrate the numbering scheme, we show in Fig. 2 the orderings of a 15×7 grid, with partitioning factor $p = 2$ and 4.

Since standard nested dissection ordering is applied locally to a number of smaller subgrids, we refer to this ordering as a *local nested dissection ordering*. It should be noted that when the grid is square, this local ordering becomes the standard nested dissection ordering.

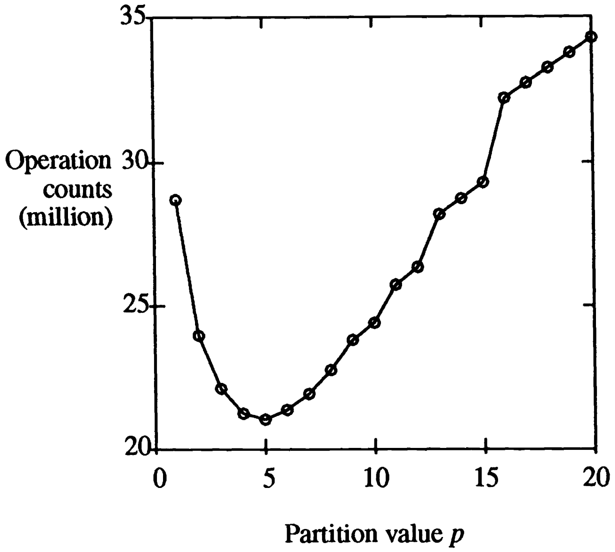


FIG. 3. *Factorization operation counts (in millions) versus partition factor for* 500*-by-*60 *grid.*

3.1. Choice of partitioning value. It is interesting to point out that the local nested dissection ordering with a partition value $p = 1$ corresponds to a hybrid nested dissection strategy introduced by Rose and Whitten [6]. We can therefore view the new scheme as a generalization of the hybrid dissection scheme of Rose and Whitten. On the other hand, for a large partition factor p close to k, the new ordering is simply the natural ordering.

In Fig. 3, we plot the number of factorization operations (in millions) against different values of p for the 500-by-60 grid problem. The minimum operation count is attained for a partition value of about 5. For comparison, we note that the standard nested dissection ordering requires 29.1 million operations to factor, while the natural ordering requires 58.4 million operations.

The observation that the operation count is minimum around $p = 5$ can be established formally for general rectangular grids. We now proceed with the analysis. The following bound on the arithmetic operations can be obtained by a straightforward count, and its tedious proof is omitted. The derivation uses complexity formulae given in [5, p. 266].

THEOREM 3.1. *The number of arithmetic operations required for factoring the sparse matrix associated with the local nested dissection ordering with partition value p is asymptotically bounded by*

$$(h-k)k^2\left(p+\frac{7}{2}+\frac{369}{12p}-\frac{125}{6p^2}\right)+\frac{829}{84}k^3.$$

COROLLARY 3.2. *The asymptotic bound in Theorem* 3.1 *is minimized at* $p \approx 4.67$.

Proof. Differentiating the coefficient expression for the term hk^2 with respect to p and equating to zero, we obtain $12p^3 - 369p + 500 = 0$. The solutions to this equation are 1.46, 4.67, and −6.13, respectively. It is easy to verify that $p \approx 4.67$ gives the minimum. □

The optimal choice of p between 4 and 5 has a very intuitive explanation. For square subgrids, we definitely want to use nested dissection. For an s-by-s subgrid, the number of nodes on its boundary is roughly $4s$. When $4s$ has reached k, it now pays to switch to a scheme similar to the natural ordering.

3.2. Computational results. In this section, we provide statistics to demonstrate the practical improvements of the local nested dissection ordering. We always use a partitioning factor p of 5. Substituting into the expression in Theorem 3.1, we obtain an asymptotic bound of $13.82hk^2 - 3.95k^3$. For a large value of h, this bound compares favorably with the bound of $18.99hk^2$ for the nested dissection scheme.

In Table 2, we tabulate statistics for various rectangular grids. We choose h-by-k grids so that the number of unknowns is $hk = 30,000$. The columns "Nonz." and "Opn. count" give the number of factor nonzeros and the number of arithmetic operations, respectively. In both columns, the numbers provided are scaled by a factor of 10^{-6}.

TABLE 2
Factorization statistics for h-by-k rectangular grids (scaled by 10^{-6}*).*

h-by-k	Natural ordering		Standard ND		Local ND	
Grids	Nonz.	Opn. count	Nonz.	Opn. count	Nonz	Opn. count
2000 × 15	0.478	4.523	0.533	7.221	0.430	4.161
1000 × 30	0.928	15.758	0.716	15.351	0.621	9.794
500 × 60	1.826	58.376	0.887	29.098	0.819	21.022
250 × 120	3.615	223.842	1.008	44.679	0.990	39.781
200 × 150	4.507	346.474	1.009	45.148	1.022	45.493

The statistics in Table 2 show that the new ordering can reduce arithmetic operations by as much as one-third over that of the nested dissection ordering. Even for the 250 × 120 problem, there is a saving of more than 10 percent. A simple extension of this approach can be applied to three-dimensional grid problems. A different partitioning factor p is obtained using the same analysis technique on such problems.

4. Concluding remarks. We have described a new ordering for rectangular grids, which combines the strengths of both the nested dissection and natural orderings. Substantial savings in factorization operations can be achieved for grids with large aspect ratios. The ordering depends crucially on the rectangular grid requirement. Nevertheless, this ordering is still useful in two ways.

First, it has been observed [7] that for a few practical sparse matrix problems, the profile-oriented reverse Cuthill–McKee ordering gives better results than the fill-reducing minimum degree ordering [5]. For example, for the problem BCSSTK16 (Engineers dam) in the Harwell–Boeing collection [2] of sparse matrices, the reverse Cuthill–McKee ordering yields a factor of 622,603 off-diagonal nonzeros, which requires 43.1 million operations to compute. However, the factor from the minimum degree ordering has 736,294 off-diagonal nonzeros, and takes 74.9 million operations. The minimum degree ordering is clearly undesirable. The performance of the local nested dissection ordering on the rectangular grids will give some insight into the possible remedy of the minimum degree ordering, so that it could be consistently better than the profile-oriented orderings in reducing fills.

Second, this new ordering on rectangular grids will be useful in the study of parallel sparse factorization algorithms. In the context of parallel elimination, the standard nested dissection ordering is desirable since its divide-and-conquer paradigm provides for independent eliminations. However, the total serial operation count for the nested dissection ordering is significantly more than that of the new ordering on rectangular grids. This gives an interesting example to study the role of fill-reducing orderings in parallel factorization.

REFERENCES

[1] I. S. Duff, A. M. Erisman, and J. K. Reid, *On George's nested dissection method*, SIAM J. Numer. Anal., 13 (1976), pp. 686–695.

[2] I. S. Duff, R. Grimes, and J. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[3] J. A. George, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 345–363.

[4] ———, *Numerical experiments using dissection methods to solve n by n grid problems*, SIAM J. Numer. Anal., 14 (1977), pp. 161–179.

[5] J. A. George and J. W. H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice Hall, Englewood Cliffs, NJ, 1981.

[6] D. J. Rose and G. F. Whitten, *A Recursive Analysis of Dissection Strategies*, in Sparse Matrix Computations, J. Bunch and D. Rose, eds., Academic Press, New York, 1976, pp. 59–84.

[7] P. Vu, personal communication, 1989.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 259–278, January 1993

021

ORDERINGS, MULTICOLORING, AND CONSISTENTLY ORDERED MATRICES*

DAVID L. HARRAR II†

Abstract. The use of multicoloring as a means for the efficient implementation of diverse iterative methods for the solution of linear systems of equations, arising from the finite difference discretization of partial differential equations, on both parallel (concurrent) and vector computers has been extensive; these include SOR-type and preconditioned conjugate gradient methods as well as smoothing procedures for use in multigrid methods. Multicolor orderings, corresponding to reorderings of the points of the discretization, often allow a local decoupling of the unknowns. Some new theory is presented which allows one to quickly verify whether or not a member of a certain class of matrices is consistently ordered (or π-consistently ordered) solely by looking at the structure of the matrix under consideration. This theory allows one to quickly ascertain that, while many well-known multicoloring schemes do give rise to coefficient matrices which are consistently ordered, many others do not. Some alternative orderings and multicoloring schemes proposed in the literature are surveyed and the theory is applied to the resulting coefficient matrices.

Key words. multicoloring, consistently ordered matrices, iterative methods, concurrent computers

AMS(MOS) subject classifications. 15, 65

1. Introduction.

The discretization by finite differences, or finite elements, of elliptic partial differential equations often leads to the solution of linear systems of equations

$$Au = f. \tag{1}$$

With the advent of parallel computers and vector processors, it has become apparent that the use of alternative orderings, i.e., other than the *natural* or *lexicographic ordering*, may increase efficiency in the implementation of many iterative methods for solving (1); these methods include the Jacobi, Gauss–Seidel, and successive overrelaxation (SOR) iterations, and various preconditioned conjugate gradient (PCG) methods, as well as smoothing procedures for use in multigrid methods.

This leads naturally to the use of the technique of *multicoloring* to decouple the unknowns at the grid points of a finite difference, or finite element, discretization of a partial differential equation. The basic idea is to "color" the grid points so that unknowns corresponding to grid points of a particular color are coupled only with unknowns of other colors. Thus all unknowns of a single color can be updated simultaneously, i.e., in parallel, or with a single vector instruction, assuming that the unknowns are stored appropriately.

In general, multicoloring with p colors corresponds to a partitioning π of the

* Received by the editors July 2, 1990; accepted for publication (in revised form) May 8, 1991. This research was supported in part by Department of Energy grant DE-FG03-89ER25073 and by the National Science Foundation under Cooperative Agreement CCR-8809615. The government has certain rights in this material.

† Department of Applied Mathematics 217-50, California Institute of Technology, Pasadena, California 91125 (dlh@ama.caltech.edu).

coefficient matrix of the system (1) into the block $p \times p$ form

$$A_p = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdot & \cdot & A_{1,p} \\ A_{2,1} & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & \cdot & \cdot & A_{p-1,p} \\ A_{p,1} & \cdot & \cdot & A_{p,p-1} & A_{p,p} \end{bmatrix}, \tag{2}$$

where the diagonal blocks A_{ii} are square. Full decoupling of all unknowns of a given color from those of other colors obtains if the A_{ii} are diagonal, and, in general, a significant number of the off-diagonal blocks A_{ij} contain only zeros. Throughout this paper we maintain the notational convention that a *single* subscript on a matrix name is used to emphasize the *block* order of that matrix; when necessary for clarity, the corresponding partitioning π carries the same subscript as in "π_p."

Multicoloring has been used ubiquitously for the solution of linear systems by iterative methods on both parallel and vector computers (for a review, see, e.g., Ortega and Voigt [16]). Although more often a multicoloring scheme is used in conjunction with SOR-type iterative methods (e.g., Adams and Ortega [2] and O'Leary [15]), it can also prove useful with PCG methods. Poole and Ortega [18] use multicoloring to carry out incomplete Cholesky preconditioning on vector computers, and Harrar and Ortega [11] used a red/black ordering to efficiently vectorize a symmetric successive overrelaxation (SSOR) preconditioner. The parallel and vector implementation of SSOR PCG (as well as SOR) via multicoloring is also discussed in Harrar and Ortega [10], where a compromise is proposed between the faster convergence rate obtained with the natural ordering and the superior degree of parallelism and/or vectorization provided by the red/black ordering (see §5.2).

When solving elliptic problems using multigrid methods, much of the computation time is spent on the relaxation procedure used at each grid level. Multicoloring is useful in this area as well. For example, Gauss–Seidel smoothing with a red/black ordering is quite effective (Foerster, Stüben, and Trottenberg [5]), alternating direction line methods are particularly robust, and *zebra* orderings (§5.1) are useful for anisotropic equations (Stüben and Trottenberg [20]).

Not long ago a fair amount of attention was given to the concept of *consistently ordered (CO) matrices* (see §2) and some generalizations thereof: generalized CO (GCO), CO(q,r) (see §6), GCO(q,r) (we note that GCO (q,r) matrices are *p-cyclic* in the sense of Varga [23]), and π-CO matrices (Young [25]). Much of the foundation of the work done in this area can be found in the classical texts, Young [25] and Varga [23]. Lately, however, interest in whether or not the coefficient matrix A of (1) is consistently ordered has somewhat waned. As a result, we often work with a system of equations that is not CO (or GCO, π-CO, etc.) when a simple permutation of the elements of A might yield a matrix with one or more of these properties. The motivation for wanting the coefficient matrix A to have one or more of these properties is discussed in §2 along with some concepts related to consistent ordering.

In §§3 and 4, we give some new theoretical results as to when matrices with a certain underlying block structure may be CO or π-CO ("block" CO). In §5, we apply these results to show the consistent ordering of some standard alternative orderings and the lack of this property for some other orderings proposed in the literature. Section 6 contains some applications of the results to another class of matrices, and in §7 we summarize our results.

2. Consistently ordered matrices and related concepts. One property that may or may not obtain for the coefficient matrix A as a result of a reordering of the unknowns is that of being a *CO matrix.* Rather than appealing directly to the definition of a CO matrix (Young [25, Def. 5.3.2]), it is often more convenient to use the notion of a *compatible ordering vector.*

DEFINITION 2.1. The vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)^T$, where the γ_i are integers, is a *compatible ordering vector* for the matrix A of order n if, for $a_{ij} \neq 0$,

$$\gamma_i - \gamma_j = \begin{cases} 1 & \text{if } i > j, \\ -1 & \text{if } i < j. \end{cases} \tag{3}$$

The usefulness of compatible ordering vectors is made clear by the following theorem.

THEOREM 2.2 (Young [25, Thm. 5.3.2]). *A matrix A is consistently ordered if and only if a compatible ordering vector exists for A.*

Determination of the optimum relaxation parameter for the SOR method applied to the system (1) via Young's classical SOR theory [25] is based upon the relationship $(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2$ between the eigenvalues μ and λ of the Jacobi and SOR iteration matrices, respectively, associated with A; ω is the relaxation parameter. The derivation of this eigenvalue relation is based upon a determinantal invariance which is true for T-matrices (see Definition 2.8 in §3). CO matrices were introduced as a more general class of matrices for which this eigenvalue relation holds. Analogous relations relating Jacobi eigenvalues to those of the corresponding SSOR iteration matrix have been obtained by Chong and Cai [3] for GCO$(k, p-k)$ matrices and Li and Varga [13] for GCO(q, r) matrices. We note also that CO matrices possess *property A* as defined by Young.

Young [24] conjectured, and Varga [21] proved, that orderings resulting in CO coefficient matrices were optimal in terms of rate of convergence for the SOR method with $\omega = 1$, i.e., the Gauss–Seidel method. However, the usefulness of this theory is not limited solely to the use of SOR-type methods. For example, Harrar and Ortega [9] used the fact that a 2-cyclic matrix is CO and a result relating the eigenvalues of the corresponding SOR and SSOR iteration matrices, to derive an optimality result for the relaxation parameter ω in the context of the m-step SSOR PCG method. We note that the effect of consistent ordering, if any, on the rate of convergence of SSOR PCG methods is not known. For more details on the motivation for desiring that a matrix be CO, see Harrar [8] and, of course, Young [25].

In the sequel, we are concerned primarily with the property of being CO for block $p \times p$ matrices of the form (2). To this end we have a weaker version of consistent ordering.

DEFINITION 2.3. Let the matrix A be partitioned as in (2) and define a $p \times p$ matrix $Z = (z_{rs})$ by

$$z_{rs} = \begin{cases} 0 & \text{if } A_{rs} = 0, \\ 1 & \text{if } A_{rs} \neq 0. \end{cases} \tag{4}$$

The matrix A is π_p-*consistently ordered* (π_p-*CO*) if Z is consistently ordered.

We note that, according to our notation, an $(n \times n)$-CO matrix is also π_n-CO.

Analogous to Definition 2.1 for a compatible ordering vector, we now introduce the concept of a π_p-*compatible ordering vector* for a block $p \times p$ matrix. This definition is intimated in Young [25]; we formalize it here.

DEFINITION 2.4. The vector $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_p)^T$, where the γ_i are integers, is a *π_p-compatible ordering vector* for the block $p \times p$ matrix A if, for Z corresponding to A, as defined in (4), (3) holds.

We state without proof the following analog of Theorem 2.2.

THEOREM 2.5. *A block $p \times p$ matrix A is π_p-CO if and only if there exists for A a π_p-compatible ordering vector.*

Obviously, a matrix of the form (2) can be π-CO and still not be CO. For example, a full 4×4 matrix partitioned as a block 2×2 matrix is π_2-CO but is not CO, since we cannot construct a compatible ordering vector for it. This is because in such a matrix we have $a_{ij} \neq 0$ for all indices i, j, and the following observation holds.

OBSERVATION 2.6. It is impossible to construct a (π_p)-compatible ordering vector for a matrix of (block) order $p > 2$, all of whose (block) elements are nonzero; that is, no (block) full matrix of (block) order $p > 2$ is (π_p)-CO.

We also note the following.

OBSERVATION 2.7. All matrices of order greater than one are π_2-CO.

As shown in Theorems 2.9 and 2.10 below, one type of matrix of the form (2) that is both π-CO and CO is a *T-matrix.*

DEFINITION 2.8. A matrix with the block tridiagonal form

$$(5) \qquad T_p = \begin{bmatrix} D_1 & H_1 & & & \\ K_1 & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & H_{p-1} \\ & & & K_{p-1} & D_p \end{bmatrix},$$

where the D_i are square diagonal matrices, is a *T_p-matrix.*

It is easily verified that $\gamma = (1, 2, \dots, p)^T$ is a π_p-compatible ordering vector for a matrix of the form (5) where the D_i need not be diagonal. Thus we have the following theorem.

THEOREM 2.9. *A block $p \times p$ tridiagonal matrix is a π_p-CO matrix.*
(This is proved in a different manner by Hageman and Young [6].) Of course, an immediate corollary is that T_p-matrices are π_p-CO.

Now, although block tridiagonal matrices are not generally CO, T-matrices are.

THEOREM 2.10 (Young [25, Thm. 5.3.1]). *A T-matrix is a CO matrix.*

The property of being π-CO is important because it is often the case that, although a given matrix may not be CO, it is π-CO for some partitioning π of A into blocks. And, in such circumstances, we may apply the results of the classical SOR theory to the partitioned matrix, i.e., we obtain an eigenvalue relation between the eigenvalues of the Jacobi and SOR iteration matrices, respectively, corresponding to the block partitioning of A. Of course, since the eigenvalues of the block Jacobi iteration matrix are a function of the matrix elements, it is generally nontrivial to compute them. But, for example, we can compute the optimal relaxation parameter ω_{opt} and the corresponding spectral radius ρ_{opt} for line SOR methods (see §5.1).

3. The addition of block bidiagonal matrices B_r to T_p. In this section we investigate what types of matrices can be added to block tridiagonal matrices (and hence also to T-matrices) so that the resulting matrix sum is still π-CO or even CO. This section is divided into three subsections. First, we examine the case in which the block tridiagonal matrix has no zero blocks on the first sub- or superdiagonal, that is, $H_i \neq 0$ or $K_i \neq 0$ for $i = 1, \dots, p-1$ in (5). Next, we consider the case in which the block $p \times p$ tridiagonal matrix has intermittent zero blocks on these diagonals, say $H_{kq} = 0$, $K_{kq} = 0$ for $k = 1, \ldots, r$ where $r = p/q$ and q is some integer that divides p evenly; the generalization to the case for which these zero blocks are spaced nonuniformly should be obvious. Finally, we investigate the case in which the block tridiagonal matrix is a T-matrix.

Note throughout this paper that all results on π-consistent ordering for block matrices imply concomitant corollaries for the case of consistent ordering of nonblock matrices as a result of the correspondence inherent in Definition 2.3. That is, all results for block $p \times p$ matrices with blocks A_{ij} in terms of π-consistent ordering imply exactly analogous results for $n \times n$ (where $n = p$) matrices with elements a_{ij} in terms of consistent ordering.

3.1. T_p has no zero blocks on the first sub- or superdiagonal. Here we consider the class of block $p \times p$ tridiagonal matrices T_p which have no zero blocks on either the first subdiagonal or the first superdiagonal. We find that, for the matrix sum $T_p + A_p$ to be π_p-CO, A_p must also be block tridiagonal. Although somewhat obvious, we state the result as a theorem in order to facilitate reference to it below. The proof illustrates the general method of proof used throughout.

THEOREM 3.1. *Let T_p be a block tridiagonal matrix of the form* (5) *where the D_i are not necessarily diagonal, and suppose that A_p has the block $p \times p$ form* (2) *where the blocks are partitioned commensurately with those of T_p. If $H_i \neq 0$ and $H_i \neq -A_{i,i+1}$, or $K_i \neq 0$ and $K_i \neq -A_{i+1,i}$, for $i = 1, \ldots, p-1$, then the matrix $T_p + A_p$ is a π_p-CO matrix if and only if $A_{ij} = 0$ if $i > j+1$ or $j > i+1$, i.e., A_p is block tridiagonal.*

Proof. Clearly the sum of block tridiagonal matrices is also block tridiagonal, so that if A_p is, then so is $T_p + A_p$. Therefore, $T_p + A_p$ is π_p-CO by Theorem 2.9.

Now, assume that $T_p + A_p$ is π_p-CO, and suppose that $A_{ij} \neq 0$ for some i, j with, without loss of generality, $j > i+1$. Since $T_p + A_p$ is π_p-CO, we can, by Theorem 2.5 construct a π_p-compatible ordering vector γ for $T_p + A_p$. Now, either $[T_p + A_p]_{i,i+1} = H_i + A_{i,i+1} \neq 0$ or $[T_p + A_p]_{i+1,i} = K_i + A_{i+1,i} \neq 0$, for $i = 1, \ldots, p-1$ since $H_i \neq -A_{i,i+1}$ or $K_i \neq -A_{i+1,i}$, respectively. Therefore, we must have $\gamma_{i+1} - \gamma_i = 1$ for $i = 1, \ldots, p-1$. Now, $[T_p + A_p]_{ij} = A_{ij} \neq 0$, so that we also require $\gamma_j - \gamma_i = 1$. However,

$$\gamma_j - \gamma_i = (\gamma_j - \gamma_{i+1}) + (\gamma_{i+1} - \gamma_i) = (\gamma_j - \gamma_{i+1}) + 1 > 1$$

since $j > i+1$, a contradiction. Therefore, we must have $A_{ij} = 0$ for $j > i+1$. The case $A_{ij} \neq 0$ with $i > j+1$ is exactly analogous. □

We note that, under the assumptions of Theorem 3.1, if T_p and A_p are T-matrices, then $T_p + A_p$ is CO by Theorem 2.10 since the sum of T-matrices is a T-matrix. However, the corresponding "only if" part of the theorem does not hold, in general, for T-matrices and consistent ordering. Although it is possible to add matrices other than T-matrices to a T-matrix to obtain a CO matrix, the only type we can add *without knowing anything about the internal structure of the off-diagonal blocks of T and A* is a T-matrix.

The following is an immediate corollary of Theorem 3.1.

COROLLARY 3.2. *Let the $n \times n$ matrix A be such that all of the elements on the first sub- or superdiagonal are nonzero. Then A is CO if and only if $a_{ij} = 0$ for $j > i+1$ and $i > j+1$.*

3.2. T_p has intermittent zero blocks on the first sub- and superdiagonal. We now consider block tridiagonal matrices that have intermittent zero blocks on the first off-diagonals. That is, we let the matrix T have the block $r \times r$ diagonal form

$$T = \operatorname{diag}(T_{11}, T_{22}, \ldots, T_{rr}), \tag{6}$$

where the $q \times q$ diagonal blocks T_{kk} are of the form

$$T_{kk} = \begin{bmatrix} A_{l+1,l+1} & A_{l+1,l+2} & & & \\ A_{l+2,l+1} & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & A_{l+q-1,l+q} \\ & & & A_{l+q,l+q-1} & A_{l+q,l+q} \end{bmatrix}, \tag{7}$$

for $k = 1, \ldots, r$, where $l = (k-1)q$ and $p = qr$. According to our notational convention, $T = T_r = T_p$, where T_r is block $r \times r$ unidiagonal with $q \times q$ blocks and T_p is block $p \times p$ tridiagonal with zero blocks every qth entry on the first sub- and superdiagonals. Since T_p is block tridiagonal, it is π_p-CO by Theorem 2.9 (it is also trivially π_r-CO). Of course, if the diagonal blocks $A_{l+i,l+i}$, $i = 1, \ldots, q$ are square diagonal matrices, i.e., $A_{l+i,l+i} = D_{l+i}$ in (7), then T_p given by (6), (7) is also a T_p-matrix and hence CO by Theorem 2.10.

Now consider the class of block $r \times r$ bidiagonal matrices of the form

$$B_r^m = \begin{bmatrix} 0 & B_{1,2}^m & & & \\ B_{2,1}^m & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & B_{r-1,r}^m \\ & & & B_{r,r-1}^m & 0 \end{bmatrix}, \tag{8}$$

where the $q \times q$ nonzero blocks $B_{k,k+1}^m$, $k = 1, \ldots, r-1$, may have only one nonzero block lower diagonal $(m = m_L)$

$$B_{k,k+1}^{m_L} = L_{k,k+1}^{m_L} = \begin{bmatrix} & & & \\ A_{l+m_L,kq+1} & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ & & & A_{kq,(k+1)q-(m_L-1)} \end{bmatrix}, \tag{9}$$

where $m_L = 2, \ldots, q$, or a nonzero block main diagonal ($m_L = 1$ in (9)), or one nonzero block upper diagonal $(m = m_U)$

$$B_{k,k+1}^{m_U} = U_{k,k+1}^{m_U} = \begin{bmatrix} A_{l+1,kq+m_U} & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ & & & A_{kq-(m_U-1),(k+1)q} \\ & & & \end{bmatrix}, \tag{10}$$

where $m_U = 2, \ldots, q$. The blocks $B_{k+1,k}^m$, $k = 1, \ldots, r-1$ have the same block structure as the blocks $(B_{k,k+1}^m)^T$. Note that T_p is block tridiagonal of block order p while B_r^m is block bidiagonal of block order $r = p/q$; the block orders are different.

We now have the following lemma.

LEMMA 3.3. *Let T_p be given by* (6), (7) *and B_r^m be given by* (8) *with* (9) *or* (10). *Then the matrix $T_p + B_r^m$ is π_p-CO for all values of*

$$m = \begin{cases} m_L = 1, \ldots, q, \\ m_U = 2, \ldots, q. \end{cases} \tag{11}$$

Proof. We show that the matrices $T_p + B_r^m$ are π_p-CO by showing the existence of π_p-compatible ordering vectors γ. We treat separately the cases in which the blocks $B_{k,k+1}^m$ have the form (9) or (10), i.e., $m = m_L$ or $m = m_U$, respectively. First, assume that $B_{k,k+1}^m = L_{k,k+1}^{m_L}$, so that they have the form (9). In this case a π_p-compatible ordering vector for $T_p + B_r^{m_L}$ is given by

$$\gamma^L = [1, \ldots, q,\ m_L + 1, \ldots, m_L + q,\ 2m_L + 1, \ldots, 2m_L + q,\ \ldots,\ (r-1)m_L + 1, \ldots, (r-1)m_L + q]^T,$$

or, in a somewhat more compact form notationally,

$$\gamma^L = [((k-1)m_L + 1, (k-1)m_L + 2, \ldots, (k-1)m_L + q), k = 1, \ldots, r]^T. \tag{12}$$

Now, suppose that $m = m_U$, i.e., $B_{k,k+1}^m = U_{k,k+1}^{m_U}$ has the form (10). Then we obtain the π_p-compatible ordering vector

$$\gamma^U = [((k-1)(m_U - 2) + 1, (k-1)(m_U - 2) + 2, \ldots, (k-1)(m_U - 2) + q), k = 1, \ldots, r]^T \tag{13}$$

for $T_p + B_r^{m_U}$. We may verify that the vectors γ^L and γ^U given by (12) and (13), respectively, are π_p-compatible ordering vectors for the matrices $T_p + B_r^m$ where the blocks $B_{k,k+1}^m$ of B_r^m have the form (9) and (10), respectively. Thus, by Theorem 2.5, $T_p + B_r^m$ is π_p-CO for all values of m given by (11). □

The regularity among the elements of the π_p-compatible ordering vectors for the matrices $T_p + B_r^m$ is quite striking. Note that the elements of these vectors corresponding to a given block $B_{k,k+1}^m$ are consecutive integers beginning with the $(k-1)q + 1$st element of the vector; this is true for $k = 1, \ldots, r$. That is, we have

$$\gamma_{(k-1)q+i} = \gamma_{(k-1)q+i-1} + 1, \quad i = 2, \ldots, q, \quad k = 1, \ldots, r. \tag{14}$$

Therefore, for a given k, the only element of γ that depends on elements corresponding to another value of k is the $(k-1)q + 1$st; the rest of the elements for that given k can be obtained using (14). This suggests that it may be possible to construct π_p-compatible ordering vectors for matrices $T_p + B_r$ (T_p given by (6), (7)) where the matrix B_r now has the somewhat more general block $r \times r$ bidiagonal form

$$B_r = \begin{bmatrix} 0 & B_{1,2}^{m_1} & & & \\ B_{2,1}^{m_1} & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & B_{r-1,r}^{m_{r-1}} \\ & & & B_{r,r-1}^{m_{r-1}} & 0 \end{bmatrix}, \tag{15}$$

where each of the $q \times q$ nonzero blocks $B_{k,k+1}^{m_k}$, $i = 1, \ldots, r-1$, has the form (9) or (10); that is, each m_k can take on any value m in (11).

LEMMA 3.4. *Let T_p be given by* (6), (7) *and let B_r be given by* (15) *with each $B^{m_k}_{k,k+1}$ given by* (9) *or* (10), *and where each m_k takes on a value of m from* (11). *Then the matrix $T_p + B_r$ is π_p-CO for all choices of B_r, i.e., for all combinations of the m_k.*

Proof. Using the π_p-compatible ordering vectors which we constructed in the proof of Lemma 3.3, we show how to construct a π_p-compatible ordering vector for $T_p + B_r$, given any B_r of the form (15). Consider the block $B^{m_k}_{k,k+1}$. If this block has the form (9) so that $m_k = m_L$ for some m_L, then we notice from (12) that in going from γ^L_{kq} to γ^L_{kq+1}, we need only add m_L to $\gamma^L_{(k-1)q+1}$. If $B^{m_k}_{k,k+1}$ has the block strictly upper triangular form (10), we see that γ^U_{kq+1} can be obtained by subtracting $m_U - 2$ from $\gamma^U_{(k-1)q+1}$. In summary, then, we have

$$\gamma_{kq+1} = \begin{cases} \gamma_{(k-1)q+1} + m_k, & \text{if } m_k = m_L, \\ \gamma_{(k-1)q+1} - (m_k - 2), & \text{if } m_k = m_U. \end{cases} \tag{16}$$

Of course, as noted above, the remaining elements are contiguous and would be calculated using (14). Now, the construction of a π_p-compatible ordering vector for T_p+B_r proceeds as follows. We take as the first q elements of the vector γ the integers $1, \ldots, q$, i.e., $\gamma_i = i, i = 1, \ldots, q$. Next, we consider the block $B^{m_1}_{1,2}$. If this block is block lower triangular, then we set $\gamma_{q+1} = \gamma_1 + m_1 = 1 + m_1$, $\gamma_{q+i} = \gamma_{q+i-1} + 1, i = 2, \ldots, q$. However, if $B^{m_1}_{1,2}$ is block strictly upper triangular, we cannot simply use (16) to calculate γ_{q+1} since for $m_1 > 2$ we would obtain a value for γ_{q+1} which was nonpositive. To mitigate this problem, we would add $m_1 - 2$ to the first q elements of γ, then calculate γ_{q+1} using (16) and the next $q-1$ elements again using (14); we can do this because, clearly from Definition 2.1, if γ is a π-compatible ordering vector, then so is $\gamma+\delta$ where δ is any constant vector. With the blocks $B^{m_k}_{k,k+1}$, $k = 2, \ldots, r-1$, we proceed in exactly the same fashion obtaining γ_{kq+1} from $\gamma_{(k-1)q+1}$ using (16) with $m = m_k$ and then using (14) to calculate the next $q - 1$ elements of γ. If $B^{m_k}_{k,k+1}$ has the form (10), we first check if $\gamma_{kq+1} = \gamma_{(k-1)q+1} - (m_k - 2) > 0$; if not, then we first add $m_k - 2$ to the thus far computed kq elements of γ. When we have proceeded through all of the blocks $B^{m_k}_{k,k+1}$, we have constructed a π_p-compatible ordering vector for $T_p + B_r$, thus, by Theorem 2.5, $T_p + B_r$ is π_p-CO. □

We now show that, if B_r is of the form (15), then T_p+B_r cannot be π_p-CO unless each of the blocks $B^{m_k}_{k,k+1}$, $k = 1, \ldots, r-1$, has one of the unidiagonal forms (9), (10).

THEOREM 3.5. *Let T_p be given by* (6), (7) *and let B_r be given by* (15). *Then the matrix sum $T_p + B_r$ is π_p-CO if and only if each $B^{m_k}_{k,k+1}$, $k = 1, \ldots, r-1$, is given by* (9) *or* (10), *where each m_k takes on a value of m from* (11).

Proof. If each $B^{m_k}_{k,k+1}$ is given by (9) or (10) with m_k taking on a value of m from (11), then $T_p + B_r$ is π_p-CO by Lemma 3.4.

We now show that if any of the blocks $B^{m_k}_{k,k+1}$ has a form different from (9) and (10), then it is impossible to construct a π_p-compatible ordering vector for $T_p + B_r$; thus the matrix sum $T_p + B_r$ is not π_p-CO by Theorem 2.5. Consider the block $B^{m_k}_{k,k+1}$ where k is now fixed and is chosen from the range of values $k = 1, \ldots, r-1$. We assert that given one nonzero block $A_{l+i,kq+j}$ (recall that $l = (k-1)q$), where $i, j \in \{1, \ldots, q\}$, in $B^{m_k}_{k,k+1}$, the only other blocks of $B^{m_k}_{k,k+1}$ which can be nonzero are those lying along the diagonal of which $A_{l+i,kq+j}$ is a member.

We treat two cases: In Case 1 ($i \geq j$), $A_{l+i,kq+j}$ is in the lower triangular portion of $B^{m_k}_{k,k+1}$ or on the main diagonal of $B^{m_k}_{k,k+1}$. In Case 2 ($i < j$), $A_{l+i,kq+j}$ is in the upper triangular portion of $B^{m_k}_{k,k+1}$.

Case 1 ($i \geq j$). From (9) we see that $B^{m_k}_{k,k+1} = B^{m_L}_{k,k+1}$ where $m_L = i - j + 1$ and $A_{l+i,kq+j} \neq 0$ for some i, j where $m_L \leq i \leq q$ and $1 \leq j \leq q - (m_L - 1)$. The parameter m_L uniquely determines the diagonal of which $A_{l+i,kq+j}$ is a member, and it can be seen that all elements of this diagonal are of the form $A_{l+i,kq+j}$ where the integer pair (i, j) is a member of the set

$$(17) \quad \Lambda_{m_L} = \{\,(i,j)|m_L \leq i \leq q, 1 \leq j \leq q - (m_L - 1) \text{ and } i - j = m_L - 1\,\}.$$

Assume now that, for some $\xi, \eta \in \{1, \ldots, q\}$, $A_{l+\xi,kq+\eta} \neq 0$ and $(\xi, \eta) \notin \Lambda_{m_L}$. Then, in order that a π_p-compatible ordering vector exist, the requirement (3) becomes

$$\gamma_{l+\xi} - \gamma_{kq+\eta} = \begin{cases} 1 & \text{if } l + \xi > kq + \eta, \\ -1 & \text{if } l + \xi < kq + \eta. \end{cases}$$

However, since $\xi, \eta \in \{1, \ldots, q\}$, we can never have $l + \xi > kq + \eta$. Noting that $kq - l = kq - (k-1)q = q$ so that $l + \xi < kq + \eta$ is always satisfied, we thus require

$$(18) \qquad \gamma_{l+\xi} - \gamma_{kq+\eta} = -1.$$

Now, since the elements of γ are consecutive for indices from $l + 1$ to kq,

$$\begin{aligned} \gamma_{l+\xi} &= (\gamma_{l+\xi} - \gamma_{l+\xi-1}) + (\gamma_{l+\xi-1} - \gamma_{l+\xi-2}) + \cdots + (\gamma_{l+2} - \gamma_{l+1}) + \gamma_{l+1} \\ (19) \qquad &= (\xi - 1) + \gamma_{l+1}. \end{aligned}$$

Similarly, the elements of γ are consecutive for indices from $kq + 1$ to $(k+1)q$ so that

(20)
$$\begin{aligned} \gamma_{kq+\eta} &= (\gamma_{kq+\eta} - \gamma_{kq+\eta-1}) + (\gamma_{kq+\eta-1} - \gamma_{kq+\eta-2}) + \cdots + (\gamma_{kq+2} - \gamma_{kq+1}) + \gamma_{kq+1} \\ &= (\eta - 1) + \gamma_{kq+1}. \end{aligned}$$

Subtracting (20) from (19) and using the first line of (16), we have

$$(21) \qquad \gamma_{l+\xi} - \gamma_{kq+\eta} = (\xi - \eta) + \gamma_{l+1} - \gamma_{kq+1} = (\xi - \eta) - m_L.$$

Substituting into (18) this gives

$$\xi - \eta = m_L - 1.$$

But then, from the definition of Λ_{m_L}, we would have $(\xi, \eta) \in \Lambda_{m_L}$, a contradiction. Therefore, we must have $A_{l+\xi,kq+\eta} = 0$ for (ξ, η) not in Λ_{m_L}.

Case 2 ($i < j$). In this case we assume that the block $B^{m_k}_{k,k+1}$ contains a nonzero block $A_{l+i,kq+j}$, where now $1 \leq i \leq q - (m_U - 1)$ and $m_U \leq j \leq q$. Thus $B^{m_k}_{k,k+1}$ has the form $B^{m_U}_{k,k+1}$ where $m_U = j - i + 1$. From (10) we see that for any element $A_{l+i,kq+j}$ of the diagonal determined by m_U the integer pair (i, j) is a member of the set

$$(22) \quad \Upsilon_{m_U} = \{\,(i,j)|1 \leq i \leq q - (m_U - 1), m_U \leq j \leq q \text{ and } j - i = m_U - 1\,\}.$$

Now, assume that $A_{l+\xi,kq+\eta} \neq 0$ for some $\xi, \eta \in \{1, \ldots, q\}$ such that $(\xi, \eta) \notin \Upsilon_{m_U}$. Then, analogous to Case 1, in order that a π_p-compatible ordering vector exist for $T_p + B_r$, we again obtain the requirement (18). Using (19) and (20) and the third line of (16), from which $\gamma_{l+1} - \gamma_{kq+1} = m_U - 2$, (21) becomes

$$\gamma_{l+\xi} - \gamma_{kq+\eta} = (\xi - \eta) + m_U - 2.$$

Substituting this into (18) we get

$$\xi - \eta \;=\; 1 - m_U,$$

so that (ξ, η) is in Υ_{m_U}, a contradiction. Thus $A_{l+\xi,kq+\eta} = 0$ for (ξ, η) not in Υ_{m_U}. Therefore, we conclude that, in order that a π_p-compatible ordering vector exist for $T_p + B_r$ so that $T_p + B_r$ is π_p-CO, the nonzero blocks $B^{m_k}_{k,k+1}$ of B_r must, for each k, have one of the unidiagonal forms (9), (10), and the proof is complete. □

3.3. T_p is a T_p-matrix. Now, we consider the case in which T_p is a T_p-matrix with intermittent zeros on the first sub- and superdiagonal. Of course, in this case, a matrix of the form $T_p + B_r$ is still π_p-CO by Theorem 3.5, but it turns out that such a matrix is also CO.

THEOREM 3.6. *Suppose that the diagonal blocks $A_{l+i,l+i}$ of* (7) *are square diagonal matrices, and let B_r be a block bidiagonal matrix of the form* (15). *Then any matrix of the form $T_p + B_r$ where T_p is given by* (6), (7) *is CO if each of the blocks $B^{m}_{k,k+1}$ of B_r is given by one of the unidiagonal forms* (9), (10).

Proof. If each $B^{m_k}_{k,k+1}$ is given by (9) or (10) with m_k taking on a value of m from (11), then we show that we can easily construct a compatible ordering vector for $T_p + B_r$ using our previous results. Let s denote the order of the diagonal blocks $A_{l+i,l+i}$. (The case in which these blocks each have different order, say s_{l+i}, is no more difficult to prove; however, the subscripting becomes overly cumbersome.) We assert that, in order to construct a compatible ordering vector γ for $T_p + B_r$, we must only take the π_p-compatible ordering vector (which we now denote γ^π) for $T_p + B_r$, constructed as in the proof of Lemma 3.4, and repeat each element s times. That is, with $l = (k-1)q$, we set

$$\gamma_{(l+i-1)s+\hat{\imath}} = \gamma^\pi_{l+i}, \quad \hat{\imath} = 1, \ldots, s; \quad i = 1, \ldots, q; \quad k = 1, \ldots, r. \tag{23}$$

In order to verify that γ constructed in this manner is a compatible ordering vector for $T_p + B_r$, we must show that (3) holds for all nonzero elements a_{gh} of $T_p + B_r$; note that $g, h \in \{1, \ldots, n\}$ where $n = qrs$. There are two situations that we must treat: $a_{gh} \neq 0$ represents an element of one of the off-diagonal blocks $A_{l+i,l+i+1}$ of (7) and $a_{gh} \neq 0$ represents an element of some $A_{l+i,kq+j}$ of $B^{m_k}_{k,k+1}$. Consider the case in which $a_{gh} \neq 0$ is an element of one of the off-diagonal blocks of (7), $A_{l+i,l+i+1}$. Then we have

$$g = (l+i-1)s + \hat{\imath}, \qquad h = (l+i)s + \hat{\jmath},$$

where $\hat{\imath}, \hat{\jmath} \in \{1, \ldots, s\}$. So, requirement (3) becomes

$$\gamma_g - \gamma_h = \gamma_{(l+i-1)s+\hat{\imath}} - \gamma_{(l+i)s+\hat{\jmath}} = \begin{cases} 1 & \text{if } (l+i-1)s + \hat{\imath} > (l+i)s + \hat{\jmath}, \\ -1 & \text{if } (l+i-1)s + \hat{\imath} < (l+i)s + \hat{\jmath}. \end{cases}$$

Since $\hat{\imath}, \hat{\jmath} \in \{1, \ldots, s\}$, we can never have $(l+i-1)s + \hat{\imath} > (l+i)s + \hat{\jmath}$, so we require that

$$\gamma_{(l+i-1)s+\hat{\imath}} - \gamma_{(l+i)s+\hat{\jmath}} = -1.$$

By (23), this is equivalent to requiring that

$$\gamma^\pi_{l+i} - \gamma^\pi_{l+i+1} = -1,$$

which is true in our construction of γ^π by (14).

Now, we consider the case in which $a_{gh} \neq 0$ is in a block $A_{l+i,kq+j}$ of $B^{m_k}_{k,k+1}$. Then we have

$$g = (l+i-1)s + \hat{\imath}, \qquad h = (kq+j-1)s + \hat{\jmath} \tag{24}$$

for some $\hat{\imath}, \hat{\jmath} \in \{1, \dots, s\}$. Proceeding as before, we obtain the requirement

$$\gamma_{(l+i-1)s+\hat{\imath}} - \gamma_{(kq+j-1)s+\hat{\jmath}} = -1.$$

By (23), we thus require that

$$\gamma^\pi_{l+i} - \gamma^\pi_{kq+j} = -1. \tag{25}$$

This is the requirement (18) with $\xi = i$, $\eta = j$, which we found to hold if and only if $(\xi, \eta) = (i, j)$ is in Λ_{m_L} or Υ_{m_U}, depending on whether $m_k = m_L$ or $m_k = m_U$, that is, if and only if $A_{l+i,kq+j}$ lies along the diagonal determined by m_k. This is true by assumption; thus (3) holds for γ_g and γ_h corresponding to a_{gh}.

Hence, for any nonzero element of either of the two "types" of nonzero blocks ($A_{l+i,l+i+1}$ and $A_{l+i,kq+j}$) of $T_p + B_r$, the corresponding elements of γ given by (23) satisfy the requirements set forth in Definition 2.1 of a compatible ordering vector. Therefore, using (23), where the elements γ^π_{kq+i} are found as in the proof of Lemma 3.4, we can construct a compatible ordering vector for $T_p + B_r$, so it is a CO matrix by Theorem 2.2. □

As was the case with Theorem 3.1 in §3.1, we cannot strengthen the above result to be an "if and only if" statement. Although the blocks $B^{m_k}_{k,k+1}$ of B_r do not necessarily have to have one of the block unidiagonal forms (9), (10) in order that the matrix sum $T_p + B_r$ (with T_p a T_p-matrix) be CO, we can only use the method of constructing a compatible ordering vector given in the proof of Theorem 3.6 if the $B^{m_k}_{k,k+1}$ do have one of these unidiagonal forms. Otherwise, we need to know something about the internal structure of the $A_{l+i,l+i+1}$ of (7) and the $A_{l+i,kq+j}$ of $B^{m_k}_{k,k+1}$.

The method of proving Theorem 3.6 can be extended in a straightforward manner to prove the following stronger and very useful result.

THEOREM 3.7. *Let the block $p \times p$ matrix A be given by* (2), *and suppose that the diagonal blocks A_{ii}, $i = 1, \dots, p$ are diagonal matrices. If A is π_p-CO, then A is CO.*

A proof of Theorem 3.7 would be similar to the following. If A is π_p-CO, then by Theorem 2.5 there exists for A a π_p-compatible ordering vector, say γ^{π_p}. Let s_i denote the order of the diagonal block A_{ii}, for $i = 1, \dots, p$. Now construct a vector γ by repeating each element $\gamma_i^{\pi_p}$ of γ^{π_p} s_i times. Then, using the method of proof used for Theorem 3.6, we would show that the vector γ, consisting of $\sum_{i=1}^p s_i$ elements, is a compatible ordering vector for A. Thus A is CO by Theorem 2.2. We note that Young intimated this result by stating: If A_p is π_p-CO, then $C_p = D_p - A_p$ is CO, where the blocks D_{ij} of D_p are given by $D_{ii} = A_{ii}$ and $D_{ij} = 0$, $i \neq j$ (Young [25, Thm. 14.3.2]).

Often, when using a multicoloring scheme, we obtain a matrix with diagonal blocks which are, in turn, diagonal. In this case, Theorem 3.7 provides an efficient way of showing whether or not that matrix is CO by simply finding a π_p-compatible ordering vector for it rather than a compatible ordering vector. In general, this should represent a substantial simplification.

We note that the following corollary of Theorem 3.7 may also be useful.

COROLLARY 3.8. *Let the block $p \times p$ matrix A have diagonal blocks A_{ii}, $i = 1, \dots, p$, which are block diagonal of block order s. Assume that A can also be partitioned as a block $t \times t$ matrix with $t = ps$. If A is π_p-CO, then A is π_t-CO.*

4. The addition of more general block matrices M_r to T_p. The π_p-compatible ordering vectors constructed in the manner prescribed in §3.2 allow for even more nonzero blocks in the matrix sum; these nonzero blocks must again take one of the unidiagonal forms (9), (10). We consider the addition of a more general class of block matrices M_r to block tridiagonal matrices T_p where the M_r have more than just two nonzero block diagonals. We consider matrices M_r of the form

$$M_r = \begin{bmatrix} 0 & M_{1,2} & \cdot & \cdot & M_{1,r} \\ M_{2,1} & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & \cdot & \cdot & M_{r-1,r} \\ M_{r,1} & \cdot & \cdot & M_{r,r-1} & 0 \end{bmatrix}, \tag{26}$$

where we again assume that this matrix is symmetrically structured. That is, the nonzero block structure of a block M_{ij} is the same as that of M_{ji}^T. In the language of previous sections, these matrices would be referred to as block "$(2r-2)$-diagonal" matrices.

We denote the off-diagonal blocks of the matrix M_r of (26) by $M_{k,k+i}$, where $k = 1,\ldots,r$ and $i = 1,\ldots,r-k$; associated with each of these blocks will be a value $m_{k,k+i}$, selected from (11), which will determine the unidiagonal structure. Thus i serves as an index for the superdiagonal (and, by the symmetrical structure assumption, the associated subdiagonal) under consideration; the case $i = 1$ was the subject of §3.

We begin with the case $i = 2$. There are several possible situations. The allowable value of $m_{k,k+2}$ depends on the values of $m_{k,k+1}$ and $m_{k+1,k+2}$; that is, $m_{k,k+2}$ depends on the values of m in the block to the left and in the block below the $k, k+2$nd block. There are four possibilities, depending on whether these values of m are of the form m_L or m_U; the corresponding allowable values of $m_{k,k+2}$ for π_p-consistent ordering are given below. (In (27), (28) by $M_{k,k+2}^{m_{k,k+2}} = 0$, we mean that $M_{k,k+2}^{m_{k,k+2}}$ has no nonzero elements.)

(i) $m_{k,k+1} = m_L = 1,\ldots,q$ and $m_{k+1,k+2} = m_L = 1,\ldots,q$,

$$\begin{aligned} m_{k,k+1} + m_{k+1,k+2} \le q &\implies m_{k,k+2} = m_L = m_{k,k+1} + m_{k+1,k+2}, \\ m_{k,k+1} + m_{k+1,k+2} > q &\implies M_{k,k+2}^{m_{k,k+2}} = 0. \end{aligned} \tag{27}$$

(ii) $m_{k,k+1} = m_U = 2,\ldots,q$ and $m_{k+1,k+2} = m_U = 2,\ldots,q$,

$$\begin{aligned} m_{k,k+1} + m_{k+1,k+2} - 2 \le q &\implies m_{k,k+2} = m_U = m_{k,k+1} + m_{k+1,k+2} - 2, \\ m_{k,k+1} + m_{k+1,k+2} - 2 > q &\implies M_{k,k+2}^{m_{k,k+2}} = 0. \end{aligned} \tag{28}$$

(iii) $m_{k,k+1} = m_L = 1,\ldots,q$ and $m_{k+1,k+2} = m_U = 2,\ldots,q$,

$$\begin{aligned} m_{k,k+1} > m_{k+1,k+2} - 2 &\implies m_{k,k+2} = m_L = m_{k,k+1} - m_{k+1,k+2} + 2, \\ m_{k,k+1} \le m_{k+1,k+2} - 2 &\implies m_{k,k+2} = m_U = m_{k+1,k+2} - m_{k,k+1}. \end{aligned} \tag{29}$$

(iv) $m_{k,k+1} = m_U = 2,\ldots,q$ and $m_{k+1,k+2} = m_L = 1,\ldots,q$,

$$\begin{aligned} m_{k+1,k+2} \le m_{k,k+1} - 2 &\implies m_{k,k+2} = m_U = m_{k,k+1} - m_{k+1,k+2}, \\ m_{k+1,k+2} > m_{k,k+1} - 2 &\implies m_{k,k+2} = m_L = m_{k+1,k+2} - m_{k,k+1} + 2. \end{aligned} \tag{30}$$

For values of i greater than 2, we use the values of any pair of m's, $m_{k,k+t}$ and $m_{k+t,k+i}$, where $0 < t < i$. For instance, to obtain the allowable value of, say $m_{k,k+4}$, we could use any of the pairs $m_{k,k+1}, m_{k+1,k+4}$ or $m_{k,k+2}, m_{k+2,k+4}$ or $m_{k,k+3}, m_{k+3,k+4}$. The formulas to be used to obtain the allowable value of $m_{k,k+i}$ are exactly (27)–(30), except that we replace $k+1$ by $k+t$ and $k+2$ by $k+i$. Note, however, that when a block was set to zero as in the second lines of (i) and (ii) above, we do not assume a value of 0 for m in that block; rather, m carries the value indicated in the corresponding first line, although greater than q.

Proceeding in this manner, it is possible to check whether a block matrix is π_p-CO. If all of the diagonal blocks $A_{l+i,l+i}$ are diagonal matrices, then, by Theorem 3.7, we can check whether the matrix is CO.

5. Application of the theory to some multicolor orderings. In this section we apply some of the results of §3 to some well-known and not so well-known orderings that appear in the literature. Although convergence properties for these orderings are beyond the scope of this paper, they can generally be found in the given references. In order to concretize some of what follows, we apply the discussion to the solution of the discretized analog of the two-dimensional Laplace problem

$$(31)\qquad \begin{aligned} \nabla^2 u = 0 \quad &\text{in} \quad \Omega = [0,1] \times [0,1], \\ u = 0 \quad &\text{on} \quad \partial\Omega. \end{aligned}$$

In the discretization we assume that there are an even number N of grid points in each direction.

5.1. Line and zebra orderings. One frequently used class of orderings is the class of *line orderings*. These are multicoloring schemes in which all of the points on a given line of the grid, or group of lines, has the same color. Thus, for example, for a two-dimensional problem on an $N \times N$ grid, a one-line ordering results in N colors, while a k-line ordering gives N/k colors (we generally choose k so that it divides N evenly).

Ordering the lines, or groups of lines, in the natural ordering from bottom to top, the coefficient matrix under the usual five-point finite difference discretization, as with a natural ordering of the grid points, has the form

$$(32)\qquad A = \text{tridiag}(-I, T, -I), \qquad T = \text{tridiag}(-1, 4 - 1).$$

(Notationally, by tridiag(A, B, C), we mean the block tridiagonal matrix with matrices B along the main diagonal, and matrices A, C along the first sub- and superdiagonals, respectively.) Here T is $N \times N$, and I is the identity matrix of order N. For $k = 1$, we have a block $N \times N$ structure. For general k, A would be partitioned as a block $\frac{N}{k} \times \frac{N}{k}$ matrix of $kN \times kN$ blocks. In each case, the matrix is block tridiagonal of block order $\frac{N}{k}$ and hence is $\pi_{N/k}$-CO by Theorem 2.9.

Next, consider a *zebra ordering*. We color all of the odd-numbered rows of the grid, say, black, and all of the even-numbered rows white. Within each color we then number the grid points in the natural ordering. Using a five-point stencil in the discretization of (31), the coefficient matrix would have the red/black (block 2×2) form

$$(33)\qquad A = \begin{bmatrix} D_1 & C \\ C^T & D_2 \end{bmatrix}, \quad D_1 = D_2 = \text{diag}(T), \quad C = \text{tridiag}(-I, -I, 0),$$

and T is given in (32). Here diag(T) is a block diagonal matrix with diagonal blocks T. D_i, C are block $\frac{N}{2} \times \frac{N}{2}$ while T, I are $N \times N$. Since A is π_2-CO with D_1, D_2 block diagonal, Corollary 3.8 indicates that the coefficient matrix for a zebra ordering is also π_N-CO.

5.2. Many-color red/black orderings. The *many-color red/black orderings* of Harrar and Ortega [10] are of two general types: row-wise red/black orderings and planar red/black orderings. The *1-row* and *2k-row red/black orderings* involve imposing a red/black ordering on every row or $2k$ rows (lines) of the grid where $k = 1, 2, \dots, N/2$ and N is the number of rows. For three-dimensional problems, we can also consider *1-plane* and *2k-plane red/black orderings* where a red/black ordering is imposed on every plane or $2k$ planes of the grid, respectively. In either case, the red and black unknowns of a given color are numbered in a natural ordering.

In general, for a $2k$-row red/black ordering the linear system is of the form

$$(34) \qquad \left[\begin{array}{ccccccc} D_1 & A_{12} & 0 & A_{14} & \cdot & \cdot & \cdot \\ A_{12}^T & D_2 & A_{23} & 0 & \cdot & \cdot & \cdot \\ 0 & A_{23}^T & D_3 & A_{34} & 0 & A_{36} & \\ A_{14}^T & 0 & A_{34}^T & D_4 & A_{45} & 0 & \\ \cdot & \cdot & 0 & A_{45}^T & \cdot & & \\ \cdot & \cdot & A_{36}^T & & & \cdot & \\ \cdot & \cdot & & & & & \cdot \end{array}\right] \left[\begin{array}{c} u_{R1} \\ u_{B1} \\ u_{R2} \\ u_{B2} \\ \cdot \\ \cdot \\ \cdot \end{array}\right] = f.$$

Here the D_i are diagonal matrices, u_{Ri} is the vector of unknowns associated with the Ri grid points, and similarly for u_{Bi}. For an $N \times N$ grid the coefficient matrix is block $\frac{N}{k} \times \frac{N}{k}$ with $kN \times kN$ blocks. For three-dimensional problems on an $N \times N \times N$ grid, a $2k$-plane red/black ordering again yields a system of the form (34), where the coefficient matrix is block $\frac{N}{k} \times \frac{N}{k}$, except that now the blocks are $kN^2 \times kN^2$ and consist of more nonzero diagonals.

The coefficient matrix of the system (34) is a matrix of the form $T_p + A_p$ where $p = N/k$; T_p is a T_p-matrix (a $T_{N/k}$-matrix), and A_p is a block $\frac{N}{k} \times \frac{N}{k}$ matrix which has all zero blocks except for $A_{i,i+3}$ (and $A_{i+3,i} = A_{i,i+3}^T$) where $i = 1, 3, \dots, N-3$. All of the blocks on the first sub- and superdiagonals of T_p are nonzero. Thus, by Theorem 3.1, $T_p + A_p$ is not π_p-CO. In order to determine whether or not the coefficient matrix of (34) is CO, we would need to investigate the internal structure of its off-diagonal blocks.

Now consider a 1-row red/black ordering. The system of equations corresponding to a five-point finite difference discretization of (31) would have the block $2N \times 2N$ form

$$(35) \qquad \left[\begin{array}{ccccccc} D_1 & A_{12} & A_{13} & 0 & \cdot & \cdot & \cdot \\ A_{12}^T & D_2 & 0 & A_{24} & \cdot & \cdot & \cdot \\ A_{13}^T & 0 & D_3 & A_{34} & A_{35} & & \\ 0 & A_{24}^T & A_{34}^T & D_4 & 0 & & \\ \cdot & \cdot & A_{35}^T & & \cdot & & \\ \cdot & \cdot & & & & \cdot & \\ \cdot & \cdot & & & & & \cdot \end{array}\right] \left[\begin{array}{c} u_{R1} \\ u_{B1} \\ u_{R2} \\ u_{B2} \\ \cdot \\ \cdot \\ \cdot \end{array}\right] = f,$$

where each block is $\frac{N}{2} \times \frac{N}{2}$. This is also the form of the system of equations for the three-dimensional analog of (31) in the case of a 1-plane red/black ordering except that each block would be $N^2/2 \times N^2/2$ and the $A_{i,i+1}$, $i = 1, \dots, N-1$ would have more nonzero diagonals.

Using the nomenclature of §3, the 1-row (1-plane) red/black coefficient matrix of (35) is a matrix of the form $T_p + B_r$, where $p = 2N$, $r = N$ ($p = 2N^2$, $r = N^2$). T_p is block tridiagonal with intermittent zero blocks every second block on the first sub- and superdiagonals. B_r is block bidiagonal where each block $B_{k,k+1}^m$ is block 2×2

with $m = m_L = 1$. This matrix is thus π_p-CO by Lemma 3.4. In fact, from the proof of Lemma 3.4, we see that a π_p-compatible ordering vector for this matrix is given by (12) with $q = 2$, that is, $\gamma^L = (1, 2, 2, 3, \ldots, 2N, 2N+1)^T$. (For the three-dimensional problem with a 1-plane red/black ordering, replace N by N^2.) Since the diagonal blocks of T_p are diagonal, Theorem 3.6 indicates that a compatible ordering vector for $T_p + B_r$ is given by

$$\gamma^L = [(1)_{N/2}, (2)_{N/2}, (2)_{N/2}, (3)_{N/2}, \ldots, (2N)_{N/2}, (2N+1)_{N/2}]^T ,$$

where $(i)_s$ denotes the s-long vector, all of whose elements are i. Thus we know that the matrix of (35) is CO *without knowing anything about the internal structure of the off-diagonal blocks.*

Returning to the $2k$-row red/black orderings, suppose that we reorder the unknowns as

$$u = (u_{R1}, u_{B1}, u_{B2}, u_{R2}, u_{R3}, u_{B3}, \ldots)^T. \qquad (36)$$

Then, the coefficient matrix has the same block form as the 1-row red/black matrix of (35), except that it is block $\frac{N}{k} \times \frac{N}{k}$ with $kN \times kN$ blocks. Thus we again obtain a π_p-CO matrix which is, in fact, also CO. Harrar and Ortega [10] discussed this reordering as a way to reduce the bandwidth of the coefficient matrix, but made no mention of CO (or π-CO) matrices.

5.3. Other orderings. In this section, we discuss some other orderings that have been proposed in the literature. The primary motivation for many of these orderings is the need for more than two colors to decouple the unknowns under a nine-point grid stencil (for a two-dimensional problem); in this case it is well known that at least four colors are necessary. Adams and Jordan [1] identified 72 distinct four-color orderings which could be used to bring about this local decoupling of unknowns. All 72 of these four-color orderings lead to matrices which are neither CO nor π_4-CO.

Adams and Jordan [1] define a *multicolor*, or *c-color*, matrix to be a block matrix of the form (2) with $p = c$ and where the diagonal blocks A_{ii} are diagonal matrices. They also define a *multicolor T-matrix* as a block tridiagonal matrix of the form

$$T_M = \text{tridiag}(U_{i-1}, M_i, L_i) , \qquad i = 1, \ldots, s \qquad (37)$$

(where, of course, U_0 and L_s do not appear in the first and last rows, respectively). In (37) the M_i are multicolor matrices, the L_i are block strictly lower triangular, and the U_i have the transposed structure of L_i, respectively. Now, by Theorem 3.7, if a multicolor matrix as defined above is π_c-CO, it is CO. However, by Theorem 3.1, if $A_{i,i+1} \neq 0$, $i = 1, \ldots, c-1$, then a multicolor matrix is not π_c-CO if any of the other A_{ij} are nonzero. If some of the $A_{i,i+1}$ do consist solely of zero entries, then Theorem 3.5 and the theory of §4 may allow one to determine whether or not the multicolor matrix is π_c-CO. However, T_M given by (37) is always π_s-CO by Theorem 2.9, and the main result of Adams and Jordan [1] is that T_M and its associated multicolor matrix have corresponding SOR iteration matrices with the same eigenvalues. Thus, if the coloring is such that the multicolor matrix has an associated multicolor T-matrix T_M, we can apply the classical SOR theory to the π_s-CO matrix T_M to gain information about the eigenvalues of the SOR iteration matrix associated with the multicolor matrix which may be neither π-CO nor CO.

Kuo and Levy [12] also consider four-color orderings for a nine-point discretization of a two-dimensional Poisson equation. Rather than analyzing the Jacobi iteration matrix in the space domain, they consider a simpler, yet equivalent, four-color iteration matrix in the frequency domain. This matrix is not π_4-CO; however, they point out that it is π_2-CO (recall Observation 2.7). Hence, they apply the SOR theory to this frequency domain matrix partitioned as a block 2×2 matrix.

O'Leary [15] considered ordering schemes that would allow for the efficient implementation on parallel computers of SOR-type iterative processes. These schemes include the "P^3," "T^3," "$H + H$," "*Cross*," and "*Box*" orderings; the names are indicative of the patterns made by the blocks of grid points of a single color, and each ordering uses three colors. These ordering schemes give rise to coefficient matrices of the form (2), where each of the diagonal blocks A_{ii} is not a diagonal matrix. For example, partitioned as a block 3×3 matrix, the coefficient matrix corresponding to a P^3 ordering of the grid points (under a nine-point stencil) has diagonal blocks A_{ii} which are block diagonal. O'Leary gives the sparsity structure for this matrix, and it is immediately apparent that it is not π_3-CO by Theorem 3.1 since blocks $A_{1,2}$ and $A_{2,3}$ contain nonzero entries but so does block $A_{1,3}$. The grid corresponding to the sparsity structure pictured in [15] has five points "per P" if the block of points P is internal to the grid. Thus many of the diagonal blocks internal to the A_{ii}, $i = 1, 2, 3$, are 5×5 and full. Therefore, this matrix is also not CO (see Observation 2.6).

Shortley and Weller [19] considered the use of $k \times k$ square blocks of points with the Gauss–Seidel method for the solution of (31). Parter and Steuerwalt [17] point out that $k \times k$ block orderings lead to coefficient matrices which satisfy block property A (Young [25] refers to this as property $A^{(\pi)}$); in fact, these orderings give rise to π-CO matrices. We obtain a coefficient matrix of the form $T_p + B_r$ where T_p is a block $(\frac{N}{k})^2 \times (\frac{N}{k})^2$ tridiagonal matrix of the form (6), (7) with zero blocks every $\frac{N}{k}$ blocks on the first sub- and superdiagonals. B_r is block $\frac{N}{k} \times \frac{N}{k}$ bidiagonal of the form (15) with nonzero blocks $B^{m_k}_{k,k+1}$ which are block $\frac{N}{k} \times \frac{N}{k}$ unidiagonal with $m_k = m_L = 1$. Thus the coefficient matrix corresponding to a $k \times k$ block ordering on an $N \times N$ grid is $\pi_{(N/k)^2}$-CO by Theorem 3.5.

Duff and Meurant [4] considered preconditioning by incomplete factorization in 17 different orderings, including the natural, red/black, zebra, and four-color orderings already discussed. The methods of this paper can be applied to many of the orderings discussed there including forward, reverse, and alternating diagonal orderings (Young showed that a forward diagonal ordering gives a CO matrix), a diagonal ordering of $k \times k$ blocks, a spiral ordering, and two block orderings attributed to Van der Vorst; see Harrar [8].

Although the Laplace problem with Dirichlet boundary conditions (BCs) yields a CO system of equations under a natural ordering, if we assume periodic BCs in either coordinate direction, the coefficient matrix is no longer CO. The theory of §4 indicates that it is also no longer π_N-CO. However, now consider a red/black ordering. With an even number N of grid points in each direction, the matrix is block 2×2 with blocks of dimension $N^2/2$. It is trivially π_2-CO (Observation 2.7), and its diagonal blocks are diagonal matrices; thus it is also CO by Theorem 3.7. We note that with an odd number of grid points in either direction, the matrix is not, in general, CO under a red/black ordering since we would no longer have full decoupling. Boundary unknowns would depend on unknowns of the same color interior to the grid so that the diagonal blocks of the coefficient matrix would no longer be diagonal. One way to overcome this difficulty may be to consider certain coloring schemes with more than two colors; see Harrar [7].

6. $T_p(q,r)$-matrices and (q,r)-CO matrices. As pointed out in §1, there are several generalizations of the class of CO matrices other than that of the class of π-CO matrices. These include *(q,r)-CO, generalized (q,r)-CO ((q,r)-GCO)*, and *π-GCO* matrices (Young [25]). In this section, we do not treat either of these "generalized" versions, although we try to give a few examples of the ways in which our previous results can be used to obtain some information concerning (q,r)-CO matrices. In particular, these results apply to the class of *$T_p(q,r)$-matrices*; this class represents a generalization of the class of T-matrices originally defined in §2. We note that in this section, q and r have no relation to the q and r of previous sections. When we mean q and r as used previously, we denote them by $\hat{q}$ and $\hat{r}$.

A formal definition of a (q,r)-CO matrix can be found in Young [25]. We note only that a $(1,1)$-CO matrix is a CO matrix in the sense of §2. Analogous to this generalization of CO matrices, we generalize the concept of a T_p-matrix to obtain Definition 6.1 (Young [25]).

DEFINITION 6.1. Let q and r be positive integers less than p. The matrix A is a *$T_p(q,r)$-matrix* if it can be partitioned into the block $p \times p$ form $A = (A_{ij})$ where, for each i, $A_{ii} = D_i$ is a square diagonal matrix and where all other blocks vanish, except possibly for the blocks $A_{i,i+r}$, $i = 1, 2, \ldots, p-r$ and $A_{i,i-q}$, $i = q+1, q+2, \ldots, p$.

Clearly a $T_p(1,1)$-matrix is a T_p-matrix as given by Definition 2.8. Also, just as T-matrices are CO, so too are $T(q,r)$-matrices (q,r)-CO. However, we try to show under what circumstances $T_p(q,r)$-matrices are also π_p-CO and, since their diagonal blocks are diagonal matrices, CO by Theorem 3.7.

Now, note that for the purposes of showing (π-)consistent ordering, a $T_p(q,r)$-matrix with either $q = 1$ or $r = 1$ can be treated as a matrix of the form $T_p + B_{\hat{r}}$ where T_p is a T_p-matrix and $B_{\hat{r}}$ is a block $\hat{r} \times \hat{r}$ bidiagonal matrix with nonzero elements on its qth or rth sub- or superdiagonal, respectively. The case $q = 1$ is of particular importance since Varga [22] gave a complete analysis of this case, and then Nichols and Fox [14] showed that the SOR method is not effective if $q > 1$. Also, the important class of p-cyclic matrices, $p \geq 2$ consists of matrices with nonzero diagonal elements which have a corresponding Jacobi iteration matrix that is permutationally similar to a $T_p(1,r)$-matrix where $r = p-1$. Therefore, in what follows, we consider only the case in which one of q and r is unity.

Appealing to Theorem 3.1, we obtain our first result.

THEOREM 6.2. *Let q and r be positive integers less than p. Suppose A is a $T_p(1,r)$-matrix such that $A_{i,i+r} \neq 0$, $i = 1, \ldots, p-r$ and $A_{i,i-1} \neq 0$, $i = 2, \ldots, p$. Similarly, suppose $\hat{A}$ is a $T_p(q,1)$-matrix such that $\hat{A}_{i,i+1} \neq 0$, $i = 1, \ldots, p-1$ and $\hat{A}_{i,i-q} \neq 0$, $i = q+1, q+2, \ldots, p$. Then A is π_p-CO and CO if and only if $r = 1$, and $\hat{A}$ is π_p-CO and CO if and only if $q = 1$.*

Proof. Let A $(\hat{A})$ be as given in the hypothesis of the theorem. Then A $(\hat{A})$ has no zero blocks on its first subdiagonal (superdiagonal). Suppose that A $(\hat{A})$ is π_p-CO. Then, by Theorem 3.1, A $(\hat{A})$ can have no zero blocks outside the first off-diagonals. That is, we must have $r = 1$ $(q = 1)$.

Now, assume that $r = 1$ $(q = 1)$. Then A $(\hat{A})$ is a $T_p(1,1)$-matrix, i.e., a T_p-matrix. Thus, by Theorem 2.9, A $(\hat{A})$ is π_p-CO, and, by Theorem 3.7, A $(\hat{A})$ is CO. □

Analogous to our progression in §3, we now consider the case in which the first subdiagonal $(q = 1)$ or the first superdiagonal $(r = 1)$ has some zero blocks. The remainder of the results of this section are stated only for the case $q = 1$, but it is trivial to adjust the proofs to handle the case $r = 1$; this should be clear from the proof of Theorem 6.2 given above.

First, we treat the case in which A is a $T_p(1,r)$-matrix, with $r = p - 1$, and has at least one zero block on the $q = 1$ subdiagonal; in this case, A is trivially a p-cyclic matrix.

THEOREM 6.3. *Let A be a $T_p(1,p-1)$-matrix, $p > 2$. A is π_p-CO and CO if and only if $A_{i,i-1} = 0$ for some $i = 2,\dots,p$.*

Proof. Let A be a π_p-CO $T_p(1,p-1)$-matrix and assume that $A_{i,i-1} \neq 0$ for $i = 2,\dots,p$. By Theorem 6.2, a $T_p(1,r)$-matrix, all of whose blocks on the first subdiagonal are nonzero, can be π_p-CO if and only if $r = 1$. However, we have $r = p - 1$, a contradiction. Thus we must have that one of $A_{i,i-1}$, $i = 2,\dots,p$ is zero. □

Now, assume that $A_{k,k-1} = 0$ where $k = 2,\dots,p$ is fixed. The nonzero blocks of A are $A_{1,p}$ and $A_{i,i-1}$, $i = 2,\dots,k-1,k+1,\dots,p$. Thus, in the construction of a π_p-compatible ordering vector for A, we require $\gamma_p - \gamma_1 = 1$ and $\gamma_i - \gamma_{i-1} = 1$, where $i = 2,\dots,k-1,k+1,\dots,p$. We may easily verify that the elements of the vector

$$\gamma^{(\pi_p)} = (p, p+1, \dots, p+k-2, k+2, k+3, \dots, p+1)^T \tag{38}$$

satisfy both of these requirements. Thus A is π_p-CO by Theorem 2.5 and CO by Theorem 3.7.

In §3.2 we considered the case in which T_p was tridiagonal with intermittent zeros every $\hat{q}$th entry on the first sub- and superdiagonal. We conclude this section with a natural extension of the results of that section. Recall that there we eventually allowed m_k, which determined the position of the sole nonzero block diagonal in the block $B^{m_k}_{k,k+1}$ of $B_{\hat{r}}$, to vary with each block. However, here we must keep m_k constant for all k so that the nonzero blocks $A_{l+i,kq+j}$ of all of the $B_{k,k+1}$ will lie along the same diagonal. This leads us to the final result of this section, which we state without proof.

THEOREM 6.4. *Let A be a $T_p(1,r)$-matrix with zero blocks every $\hat{q}$th position on the first ($q = 1$) subdiagonal. Suppose that $r = \hat{q} + m_U$ for some $m_U = 2,\dots,\hat{q}$ or $r = \hat{q} - (m_L - 2)$ for some $m_L = 1,\dots,\hat{q}$. If the rth superdiagonal has $m - 1$ zero blocks following every $\hat{q} - (m-1)$ possibly nonzero blocks (where $m = m_U$ or $m = m_L$, depending on whether $r = r(m_U)$ or $r = r(m_L)$, respectively), then A is π_p-CO and CO.*

7. Summary.

Multicoloring provides a valuable technique to increase efficiency in implementing many iterative processes (SOR-type, PCG, multigrid, etc.) to solve linear systems of equations, especially on today's parallel and vector computers.

Application of Young's classical SOR theory is valid when the coefficient matrix of the system to be solved is CO. Often, especially when a multicoloring scheme is introduced, we obtain a coefficient matrix that is not CO; however, this matrix may be π-CO (block CO) for some partitioning π, although determination of whether or not a given matrix is CO or π-CO is generally nontrivial. Though computer programs exist to determine consistent ordering (Young [25]), these may be impractical for very large matrices and do not, in general, take into account the sparsity structure inherent in the coefficient matrices corresponding to multicolored systems. We have presented some theory which allows us to ascertain quickly whether matrices which have an underlying block tridiagonal structure are (π-)CO or not; such matrices are often obtained when a multicoloring scheme is used.

We applied the theory to ordering schemes from the literature to show that while some commonly used orderings give rise to CO or π_p-CO ($p > 2$) matrices, many others do not. This is particulary true for multicolor orderings with more than two colors.

Acknowledgments. The author would like to thank Professors Herbert B. Keller, James M. Ortega, and David M. Young for looking over preliminary versions of this manuscript and Douglas A. Waetjen for helpful discussions. The detailed comments of two of the referees (one of whom pointed out the similarity between Theorem 3.7 and Theorem 14.3.2 of Young [25]) are also greatly appreciated.

REFERENCES

[1] L. ADAMS AND H. JORDAN, *Is* SOR *color-blind?*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 490–506.

[2] L. ADAMS AND J. ORTEGA, *A multi-color* SOR *method for parallel computation*, in Proc. 1982 Internat. Conf. Parallel Processing, Bellaire, MI, 1982, pp. 53–58.

[3] L. CHONG AND D.-Y. CAI, *Relationship between eigenvalues of Jacobi and* SSOR *iteration matrices for weakly p-cyclic matrix*, J. Comput. Math. Coll. Univ., 1 (1985), pp. 79–84. (In Chinese.)

[4] I. DUFF AND G. MEURANT, *The effect of ordering on preconditioned conjugate gradients*, BIT, 29 (1989), pp. 635–658.

[5] H. FOERSTER, K. STÜBEN, AND U. TROTTENBERG, *Non-standard multigrid techniques using checkered relaxation and intermediate grids*, in Elliptic Problem Solvers, M. Schultz, ed., Academic Press, New York, 1981, pp. 285–300.

[6] L. HAGEMAN AND D. YOUNG, *Applied Iterative Methods*, Academic Press, New York, 1981.

[7] D. HARRAR II, *Multicolor orderings for the concurrent iterative solution of non-Dirichlet problems*, manuscript.

[8] ———, *Alternative orderings, multicoloring schemes, and consistently ordered matrices*, Tech. Rep. CRPC-90-8, Dept. of Applied Mathematics, California Institute of Technology, Pasadena, CA, 1990.

[9] D. HARRAR II AND J. ORTEGA, *Optimum m-step* SSOR *preconditioning*, J. Comput. Appl. Math., 24 (1988), pp. 195–198.

[10] ———, *Multicoloring with lots of colors*, in Proc. Third Internat. Conf. Supercomput., Crete, Greece, 1989, pp. 1–6.

[11] ———, *Solution of three-dimensional generalized Poisson equations on vector computers*, in Iterative Methods for Large Linear Systems, Academic Press, New York, 1990.

[12] C.-C. KUO AND B. LEVY, *A two-level four-color* SOR *method*, SIAM J. Numer. Anal., 26 (1989), pp. 129–151.

[13] X. LI AND R. VARGA, *A note on the* SSOR *and* USSOR *iterative methods applied to p-cyclic matrices*, in Iterative Methods for Large Linear Systems, Academic Press, New York, 1990.

[14] N. NICHOLS AND L. FOX, *Generalized consistent ordering and optimum successive overrelaxation factor*, Numer. Math., 13 (1969), pp. 425–433.

[15] D. O'LEARY, *Ordering schemes for parallel processing of certain mesh problems*, SIAM J. Sci. Statist. Comput., 6 (1984), pp. 761–770.

[16] J. ORTEGA AND R. VOIGT, *Solution of partial differential equations on vector and parallel computers*, SIAM Rev., 27 (1985), pp. 149–240.

[17] S. PARTER AND M. STEUERWALT, *On K-line and $K \times K$ block iterative schemes for a problem arising in three-dimensional elliptic difference equations*, SIAM J. Numer. Anal., 17 (1980), pp. 823–839.

[18] E. POOLE AND J. ORTEGA, *Multicolor* ICCG *methods for vector computers*, SIAM J. Numer. Anal., 24 (1987), pp. 1394–1418.

[19] G. SHORTLEY AND R. WELLER, *The numerical solution of Laplace's equation*, J. Appl. Phys., 4 (1938), pp. 334–348.

[20] K. STÜBEN AND U. TROTTENBERG, *Multigrid methods: Fundamental algorithms, model problem analysis, and applications*, in Multigid Methods, Proceedings, Koln–Porz, 1981, W. Hackbush and U. Trottenberg, eds., Springer–Verlag, Berlin, New York, 1982, pp. 1–176.

[21] R. VARGA, *Orderings of the successive overrelaxation scheme*, Pacific J. Math., 9 (1959), pp. 925–939.

[22] ———, *p-cyclic matrices: A generalization of the Young–Frankel successive overrelaxation scheme*, Pacific J. Math., 9 (1959), pp. 617–628.

[23] R. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.

[24] D. Young, *Iterative Methods for Solving Partial Differential Equations of Elliptic Type*, Ph.D. thesis, Harvard University, Cambridge, MA, 1950.

[25] ———, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 1, pp. 279–316, January 1993

022

FORWARD INSTABILITY OF TRIDIAGONAL QR*

BERESFORD N. PARLETT[†] AND JIAN LE[‡]

Abstract. The QR algorithm is the standard method for finding all the eigenvalues of a symmetric tridiagonal matrix. It produces a sequence of similar tridiagonals. It is well known that the QR transformation from T to $\hat{T}$ is backward stable. That means that the computed $\hat{T}$ is exactly orthogonally similar to a matrix close to T. It is also known that sometimes the computed $\hat{T}$ is not close to the exact $\hat{T}$. This is caused by the occasional extreme sensitivity of $\hat{T}$ to changes in T or the shift, and will be referred to as forward instability of the (computed) QR algorithm.

For the purpose of computing eigenvalues the property of backward stability is all that is required. However, the QR transformation has other uses and then forward stability is needed.

This paper gives examples, analyzes the forward instability, and shows that it occurs only when the shift causes "premature deflation." It is shown that forward stability is governed by the size of the last entry in normalized eigenvectors of leading principal submatrices, and the extreme values of the derivative of each entry in $\hat{T}$ as a function of the shift are found.

Key words. QR transformation, tridiagonal, sensitivity

AMS(MOS) subject classification. 65F15

1. Summary.

The QR transformation is a complicated similarity transformation that depends on a parameter σ, called the shift, and that preserves the tridiagonal form of real symmetric matrices. A QR algorithm consists of a strategy for the choice of σ and the associated sequence of QR transformations. Wilkinson's shift strategy (see Example 2.4) produces a sequence of shifts that always converges to an eigenvalue. This algorithm is implemented in several routines in the EISPACK and LAPACK libraries. See [1], [4], [5], and [8].

Each QR transform of an $n \times n$ real symmetric tridiagonal matrix T may be computed by a sequence of $n-1$ similarity transformations using plane rotations. In [8], Wilkinson showed that any algorithm employing a limited sequence of orthogonal similarity transformations is backward stable in finite precision arithmetic. This means that the output matrix is exactly orthogonally similar to a perturbation of the initial matrix that is tiny in norm.

For the purpose of computing eigenvalues, backward stability is completely satisfactory. However, the QR transform has other uses as well. It occurs in inverse eigenvalue problems (see [2]) and as a deflation procedure in contexts such as the Lanczos algorithm. In these applications forward stability is needed. That means that the computed output should be close, in norm, to the output in exact arithmetic.

Unfortunately, the computed QR transformation sometimes exhibits violent forward instability. In this paper we show examples of this phenomenon and analyze it for real symmetric tridiagonal matrices. Definition of the QR transformation $\hat{T}$ of the pair T, σ is given at the beginning of §2.1 and the sequence of plane rotations is shown in (2.18). In the rest of this section we summarize our findings for the knowledgeable reader. We follow Householder conventions in notation with a few exceptions.

* Received by the editors August 1, 1988; accepted for publication (in revised form) September 5, 1991. This work was supported in part by Office of Naval Research contract N00014-90-J-1372.

[†] Department of Mathematics and Computer Science Division, Electrical Engineering and Computer Science Department, University of California, Berkeley, California 94720 (parlett@math.berkeley.edu).

[‡] EEG Systems Laboratory, San Francisco, California 94107.

1. Forward instability can appear in exact arithmetic. It is merely the extreme sensitivity of some entries of the transform $\hat{T}$ to small changes in T and/or σ. Roundoff error is not needed to provoke this phenomenon. As a thought experiment we can consider an unreduced tridiagonal T with an eigenvalue λ that is irrational. If σ is a very accurate rational approximation to λ, then it is possible for the exact QR transforms of T with respect to λ and with respect to σ, to differ, in certain entries, in all figures. For this reason we need only look at the sensitivity of $\hat{T}$ as a function of σ and ignore implementation details such as whether implicit or explicit shifting is used.

2. The QR transformation is defined for all square matrices. Its instability, when it occurs, is inherited from the sensitivity of the Q factor of a matrix ($B = QR$) and this sensitivity can be bounded by traditional perturbation theory. This was done in [7] and further extension and refinement of that work is in progress.

However, for real symmetric unreduced tridiagonal matrices T, the situation is simpler and we have been able to replace perturbation bounds by exact derivatives of the transform $\hat{T}$ with respect to the shift σ.

To explain our results we need the last entries in certain eigenvectors. Let T_k denote the leading $k \times k$ submatrix of T. Let $\lambda_i^{(k)}, i = 1, \ldots, k$ denote the spectrum of T_k and let ω_{ik} or $\omega_{i,k}$ (when the comma is needed) denote the magnitude of the last entry of the normalized eigenvector for $\lambda_i^{(k)}$. We chose the last letter of the Greek alphabet to remind us that these are the last entries.

Our main result is that forward instability occurs if and only if σ is very close to a $\lambda_i^{(k)}$ with a tiny ω_{ik}. More precisely, there are entries of $\hat{T}$ whose derivatives with respect to σ are $O(1/\omega_{ik})$ when σ is close to $\lambda_i^{(k)}$. This is the only way large values can occur in the derivatives. The details are in Theorem 4.3. We recall that the derivatives are not proportional to $\|T\|$. The effect of varying sizes in the off-diagonal entries β_i of T is hidden in the values of the $\{\omega_{ik}\}$.

One simple result (see proof of Theorem 4.3(i)) is that, for the last off-diagonal entry $\hat{\beta}_n$ of $\hat{T}$,

$$\frac{d}{d\sigma}\hat{\beta}_n(\sigma)\big|_{\sigma=\lambda_i^{(n)}} = (-1)^i \tan\theta_n,$$

where θ_n is the last rotation angle in the QR factorization of $T - \sigma I$ and

$$|\cos\theta_n| = \omega_{in}.$$

Figure 1 shows that $\hat{\beta}_n'$ is only interesting for σ very close to an eigenvalue.

3. Forward instability, if it occurs, is always preceded by what we call *premature deflation of* σ. As mentioned in the second paragraph, $\hat{T}$ may be computed by a sequence of plane rotations. Let $T^{(1)} = T$, $T^{(i)} = \Theta_i T^{(i-1)} \Theta_i^t$, $i = 2, \ldots, n$, $\hat{T} = T^{(n)}$. The intermediate matrices $T^{(k)}$ that occur in the transformation of $T^{(1)}$ into $T^{(n)}$ (see (2.18)) depart from tridiagonal form only in positions $(k-1, k+1)$ and $(k+1, k-1)$. If the entries $(k, k-1)$ and $(k, k+1)$ of $T^{(k)}$ are sufficiently small and the (k, k) entry equals σ to working accuracy, then row and column k could be deleted from $T^{(k)}$ with little change to the remaining eigenvalues. We say that σ has been deflated from T at minor step k instead of at the expected place $T^{(n)} = \hat{T}$. Most implementations do not look for this phenomenon and so do not see it. In such cases the computed version of $\hat{T}$ below the kth row bears little resemblance to the output in exact arithmetic. In particular, deflation at minor step n will not occur despite the fact that σ is an eigenvalue to working precision. This is discussed in §5 and shown in Example 2.1.

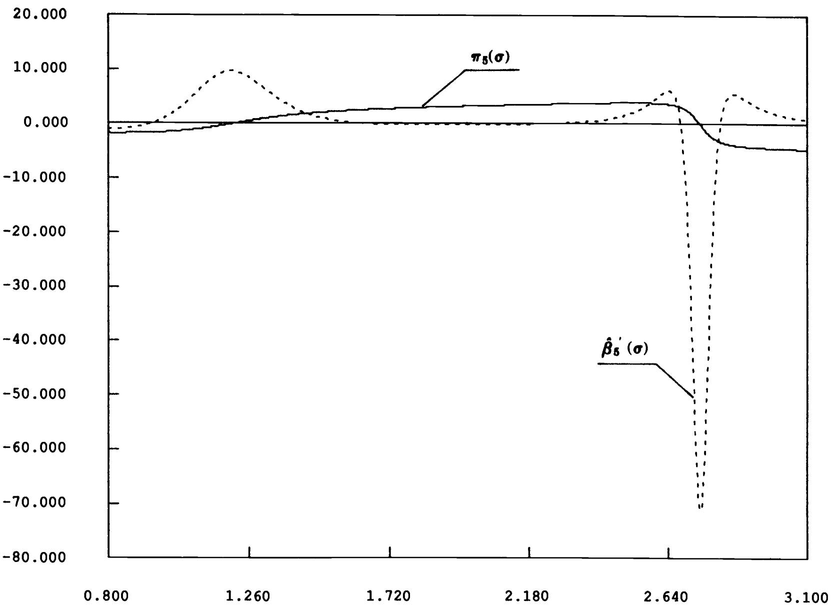


FIG. 1. *$\hat{\beta}_5'(\sigma)$ and $\pi_5(\sigma)$ on* [0.8, 3.1] *with $T = W_5^-$, defined in Example* 2.3.

4. A standard QR algorithm is somewhat protected from suffering from forward instability because of the preceding steps (transforms) in the algorithm. By the step at which σ is an eigenvalue, the last off-diagonal β_n is small and the eigenvector for σ usually has a large value for ω_{in}. This protection is not complete and Example 2.4 shows instability that occurs when Wilkinson's shift is used. However, our analysis shows that a shift strategy that picks an eigenvalue far from α_n is bound to incur forward instability when β_n is small. We are inclined to say that such a strategy has picked the "wrong" eigenvalue.

5. If $\beta_{k+1} = 0$, $k + 1 < n$, and if $\sigma = \lambda_i^{(k)}$ then the kth column of Q is not uniquely defined. Hence forward instability is inevitable close to this situation. This is when washout of the shifts occurs in the implicit implementation of QR (see [7]). This occurrence of local forward instability does not perturb the rest of the matrix when the explicit shift is used. We do not pursue this aspect further in this paper. From our point of view, when $\sigma = \lambda_i^{(k)}$ and $\beta_{k+1} = 0$, then the deflation that occurs at step k, though premature for T_n, is not premature for T_k.

6. Ultimate shifts. In [4] it was suggested that an efficient strategy for computing the spectral factorization of T would use two phases. First find the eigenvalues by any means. The QR algorithm without accumulation of the plane rotations is a leading candidate for the job. Second, run the QR algorithm with accumulation of the rotations, but using the eigenvalues as shifts. In this way the number of transformations to be accumulated is minimized. Our analysis in this paper shows that this strategy is likely to encounter *forward instability* and this might detract from the accuracy of the computed eigenvectors. However, if the monitoring algorithm described in §5 is used to terminate a QR transform at premature deflation, then this

danger may be greatly reduced. More investigation is needed.

7. Deflation in the Lanczos algorithm. A good way to compute a few eigenpairs for a large sparse symmetric matrix is to use the Lanczos algorithm; see, for example, [1] and [4]. The algorithm gradually builds up a tridiagonal matrix adding a row and a column at each step. After k steps some of the eigenvalues of the $k \times k$ tridiagonal T_k "settle down" and change very little as the algorithm proceeds. It would be convenient to deflate a converged eigenvalue from T_k. However, our analysis shows that this must be done at exactly the right step. After that, forward instability will occur. In such an application the monitoring scheme presented in §5 must be used. Indeed it was the failure to deflate eigenvalues of full accuracy that led us to this study of forward instability.

To a reader inclined to complain that we treat only instability due to a variation in the shift σ, we say two things. First, there is little loss in generality since the ω_{ik} are continuous functions of the entries of an unreduced T. A perturbed T, whatever the cause of the perturbation, will have a new set of $\lambda_i^{(k)}$ and ω_{ik}. Great sensitivity in transforming the new T will occur if σ is close to any of the new $\lambda_i^{(k)}$ with new ω_{ik} that are tiny. Second, we mention that this approach encompasses nonsymmetric Hessenberg matrices and arbitrary perturbations. Essentially, the condition for forward instability is the same. A tiny value of ω_{ik} signals that the first k columns of $T - \lambda_i^{(k)} I$ are almost linearly dependent. Note that if the first k columns of a matrix form a linearly dependent set, then the QR factorization process loses uniqueness at step k unless $k = n$.

Here is the plan of the paper. Section 2 presents notation and basic results on QR and §2.3 gives a set of examples of forward instability. Section 3 says more about the QR intermediate quantity π_k (shown in (2.1) and defined in (2.13)) than the reader will want to know. It is worth mentioning that all entries in $\hat{T}$ can be expressed in terms of the π_k and the entries in T. Not only do we give bounds on the value of π_k and its derivatives, but we give a simple model for it, and related functions, near each $\lambda_i^{(k)}$. The results are used in §4 and so §3 could be skipped by readers who have a little faith. Section 4 presents $\hat{T}'$, the derivative with respect to σ of the QR transform $\hat{T}$. Upper bounds are given first (Theorem 4.1) and show that huge values are only possible when σ is close to an eigenvalue of some principal submatrix T_k. Theorem 4.2 shows that the derivatives at the special points $\lambda_i^{(k)}$ can only be huge if σ is close to the spectrum of two successive principal submatrices. However, these results only give necessary conditions for instability. To show that small ω_{ik} are also sufficient we could see no other way than to prove Theorem 4.3. Again there is more detail than the reader may care for but the theorem is needed to establish the constants behind the O's and so to establish that values like $1/\omega_{ik}$ are attained. The functions $\hat{\beta}_k'$ can exhibit quite violent behavior close to some $\lambda_i^{(k)}$. For some of our examples the models are valid on intervals of width 10^{-14} and certain spectral points differ by 10^{-28}. A simple version of Theorem 4.3 is given in the preamble just before the detailed proof. All the results come from applying the linear model for π_k established in §3. Section 5 shows how ω_{ik} occurs in a lower bound on the premature deflation indicator. If neither β_{k+1} nor ω_{ik} is small then premature deflation cannot occur at that step.

Finally, we mention the figures whose contemplation, at the right moment, should help greatly in explaining the model and its application in Theorem 4.3. We chose a tame situation for Figs. 2 and 3 to illustrate typical situations and yet to avoid drastic rescaling of the y-axis.

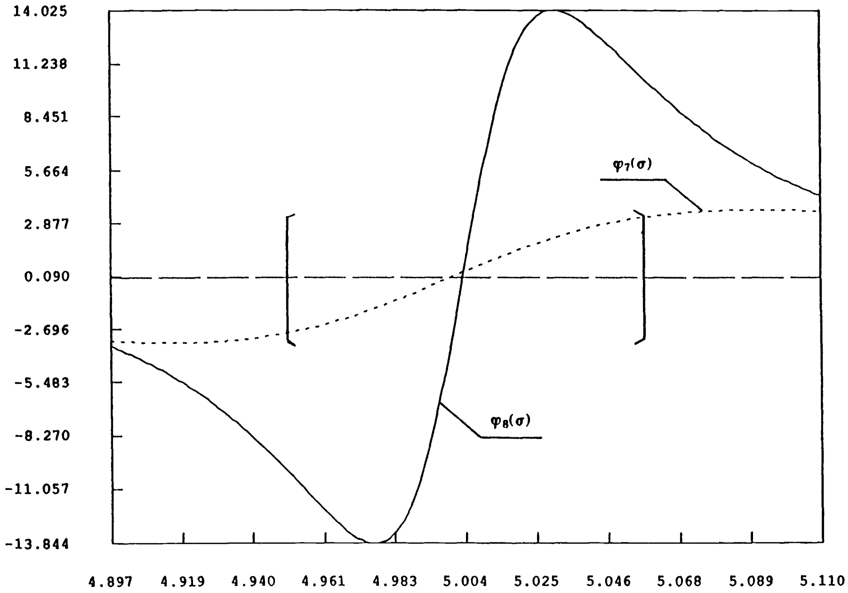


FIG. 2. *φ_8 and φ_7 on $[\lambda_5^{(8)} - \omega_{5,8}m_8(5), \lambda_5^{(8)} + \omega_{5,8}m_8(5)]$ with $T = W_{17}^-$ (see* §3.2).

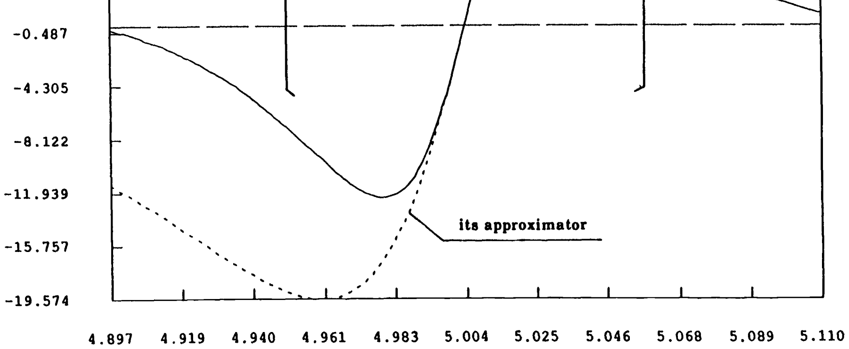


FIG. 3. *β_8' and its approximator defined in* (4.15) *on $[\lambda_5^{(8)} - \omega_{5,8}m_8(5), \lambda_5^{(8)} + \omega_{5,8}m_8(5)]$ with $T = W_{17}^-$ (see* §3.2).

2. QR basics and examples.

2.1. Relationships. For any real tridiagonal matrix T and any scalar σ (called the shift), the associated QR transformation $T \to \hat{T}$ is defined as follows:

$$T - \sigma I = QR, \qquad \hat{T} = RQ + \sigma I = Q^t T Q.$$

Let

$$T = T_n := \operatorname{tridiag}\begin{pmatrix} & \beta_2 & \beta_3 \cdots \beta_n & \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ & \beta_2 & \beta_3 \cdots \beta_n & \end{pmatrix}, \quad \beta_i > 0 \quad (i = 2, \ldots, n),$$

and let σ be the shift. T is said to be *unreduced* when $\beta_i \neq 0$, $i = 2, 3, \ldots, n$. Without loss of generality we may then assume that $\beta_i > 0$, $i = 2, 3, \ldots, n$. The QR factorization of $T - \sigma I$ may be represented by means of a sequence of plane rotations $\Theta_k, k = 2, \ldots, n$. Here Θ_k differs from the identity only in rows and columns $k - 1$ and k where it has entries

$$\begin{pmatrix} c_k & s_k \\ -s_k & c_k \end{pmatrix},$$

and $c_k = \cos\theta_k, s_k = \sin\theta_k$. We let the dimension of Θ_k be given by its context; sometimes $k \times k$, sometimes $n \times n$. The angle θ_k is chosen to annihilate the entry in the $(k, k-1)$ position of the matrix $\Theta_{(k-1)}\Theta_{(k-2)} \cdots \Theta_2(T - \sigma I)$ and to lie in $[0, \pi]$. The result of this stage is written as

$$\Theta_k \cdots \Theta_2(T - \sigma I) = \begin{bmatrix} \xi_1 & \zeta_1 & s_2\beta_3 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \xi_{k-1} & \zeta_{k-1} & s_k\beta_{k+1} \\ & & & & \pi_k & c_k\beta_{k+1} \\ & & & & \beta_{k+1} & \alpha_{k+1} - \sigma & \cdot \\ & & & & & \cdot & \cdot \end{bmatrix}. \tag{2.1}$$

The first $(k-1)$ rows are in "final form" as entries of R, while the transitory but important row k will be overwritten at the next rotation. The last row of R is

$$(0, 0, \ldots, 0, \pi_n) \equiv \pi_n e_n^t,$$

where e_k denotes the kth column of the identity matrix and v^t is the transpose of v.

The Q factor is upper Hessenberg as well as orthogonal and is specified by the $(n-1)$ rotation angles $\theta_2, \ldots, \theta_n$. The detailed structure of Q is:

$$Q = \Theta_2^t \cdots \Theta_n^t = \begin{bmatrix} c_2 & (-s_2)c_3 & (-s_2)(-s_3)c_4 & \cdot & \cdot & (-s_2)\cdots(-s_n) \\ s_2 & c_2c_3 & c_2(-s_3)c_4 & \cdot & \cdot & \cdot \\ & s_3 & c_3c_4 & & \cdot & \cdot \\ & & s_4 & \cdot & \cdot & \cdot \\ & & & \cdot & c_{n-1}c_n & c_{n-1}(-s_n) \\ & & & & s_n & c_n \end{bmatrix}.$$

The sequence $\{c_2, c_3, \ldots, c_n\}$ effectively defines Q and the elements are often called its Schur parameters; see [2].

We allow the (n, n) entry of R, namely π_n, to have the same sign as $\det[T - \sigma I]$, and $\det[Q]$ equals one in our presentation. This is a trivial departure from the convention that the diagonal of R be nonnegative.

The last column of the $k \times k$ matrix $\Theta_2^t \cdots \Theta_k^t$ plays an important role and is given a name:

$$y_k = \begin{bmatrix} (-s_2)\cdots(-s_k) \\ c_2(-s_3)\cdots(-s_k) \\ \cdot \\ \cdot \\ \cdot \\ c_{k-1}(-s_k) \\ c_k \end{bmatrix}. \tag{2.2}$$

Note that $y_n = Qe_n = q_n$. For $k < n$ we let $y_k^{(n)}$ denote a vector of the form

$$\begin{pmatrix} y_k \\ 0 \end{pmatrix} \in R^n.$$

The following relations are used frequently in the rest of the paper.

LEMMA 2.1. *With the notation developed above,*

$$Ty_k^{(n)} - y_k^{(n)}\sigma = \pi_k e_k + c_k\beta_{k+1}e_{k+1}, \qquad k < n, \tag{2.3}$$

$$\|Ty_k^{(n)} - y_k^{(n)}\sigma\|^2 = \pi_k^2 + c_k^2\beta_{k+1}^2, \qquad k < n, \tag{2.4}$$

$$\left(y_k^{(n)}\right)^t \left(Ty_k^{(n)} - y_k^{(n)}\sigma\right) = \pi_k c_k = \left(y_k^{(n)}\right)^t Ty_k^{(n)} - \sigma, \qquad k \le n, \tag{2.5}$$

$$(T - \sigma I)y_n^{(n)} = \pi_n e_n. \tag{2.6}$$

Proof. Equate the kth row on each side of (2.1) and transpose to get

$$(T - \sigma I)\Theta_2^t \cdots \Theta_k^t e_k = Ty_k^{(n)} - y_k^{(n)}\sigma = \pi_k e_k + c_k\beta_{k+1}e_{k+1}.$$

Since (2.1) holds for $k < n$, (2.3) is true for $k < n$. Equations (2.4) and (2.5) are the direct results of (2.3). Equation (2.6) is a special case of (2.3) since $\beta_{n+1} = 0$ when $k = n$. □

Equally important is the relation between $T_k \in R^{k\times k}$ and y_k.

LEMMA 2.2. *With the notation developed above,*

$$T_k y_k - y_k\sigma = \pi_k e_k, \qquad 1 \le k \le n, \tag{2.7}$$

$$y_k^t T_k y_k - \sigma = \pi_k c_k, \qquad 1 \le k \le n. \tag{2.8}$$

Proof. Equate the first k rows of (2.3) to get (2.7) for $k < n$. Equation (2.6) covers the case for (2.7) when $k = n$. Equation (2.8) is the direct result of (2.2). □

In (2.1) the relations between ξ_k, c_k, s_k, and π_k are $c_1 = 1$, $\pi_1 = \alpha_1 - \sigma$, and for $k = 2, \ldots, n$,

$$\xi_{k-1} = \sqrt{\pi_{k-1}^2 + \beta_k^2}, \tag{2.9}$$

$$c_k = \pi_{k-1}/\xi_{k-1}, \tag{2.10}$$

$$s_k = \beta_k/\xi_{k-1}, \tag{2.11}$$

$$\zeta_{k-1} = (c_k, s_k)\begin{pmatrix} \beta_k c_{k-1} \\ \alpha_k - \sigma \end{pmatrix} = c_{k-1}c_k\beta_k + s_k(\alpha_k - \sigma), \tag{2.12}$$

$$\pi_k = (-s_k, c_k)\begin{pmatrix} \beta_k c_{k-1} \\ \alpha_k - \sigma \end{pmatrix} = -s_k c_{k-1}\beta_k + c_k(\alpha_k - \sigma). \tag{2.13}$$

Finally, for $\hat{T} = RQ + \sigma I$, we have

$$\hat{\beta}_k = \xi_k s_k, \tag{2.14}$$

$$\hat{\alpha}_k = c_k\pi_k - c_{k+1}\pi_{k+1} + \alpha_{k+1}. \tag{2.15}$$

The last result comes from the invariance of the trace in $\Theta_k \cdots \Theta_2 T \Theta_2^t \cdots \Theta_k^t$ (see (2.17) below)

$$c_k\pi_k + \alpha_{k+1} - \sigma = \hat{\alpha}_k - \sigma + c_{k+1}\pi_{k+1}. \tag{2.16}$$

When $k = n$,

$$\hat{\beta}_n = \pi_n s_n, \qquad \hat{\alpha}_n = c_n\pi_n + \sigma.$$

For future reference we show the active part of the matrix $\Theta_k \cdots \Theta_2 T \Theta_k^t \cdots \Theta_k^t$ and the bulge in positions $(k+1, k-1)$ and $(k-1, k+1)$.

$$\begin{array}{ccccc} \cdot & \hat{\beta}_{k-1} & & & \\ \hat{\beta}_{k-1} & \hat{\alpha}_{k-1} & \pi_k s_k & s_k\beta_{k+1} & \\ & \pi_k s_k & c_k\pi_k + \sigma & c_k\beta_{k+1} & \\ & s_k\beta_{k+1} & c_k\beta_{k+1} & \alpha_{k+1} & \beta_{k+2} \\ & & & \beta_{k+2} & \cdot \end{array} \tag{2.17}$$

In a study of the QR transform it is useful to exploit two different interpretations of $\hat{T} = \Theta_n \cdots \Theta_2 T \Theta_2^t \cdots \Theta_n^t$. First interpretation:

$$\begin{array}{rcl} T & \rightarrow & R = \Theta_n \cdots \Theta_2 T \\ & \rightarrow & \hat{T} = R\Theta_2^t \cdots \Theta_n^t. \end{array}$$

Second interpretation:

$$\begin{array}{rcl} T & \rightarrow & T^{(2)} = \Theta_2 T \Theta_2^t \\ & \rightarrow & T^{(3)} = \Theta_3 T^{(2)} \Theta_3^t \\ & \rightarrow & T^{(n)} = \Theta_n T^{(n-1)} \Theta_n^t. \end{array} \tag{2.18}$$

Only T and $T^{(n)} = \hat{T}$ are tridiagonal matrices; the intermediate $T^{(k)}$ contain the bulge.

2.2. Deflation. The assumption that T is unreduced guarantees, in exact arithmetic, that the spectra of T_k and its submatrix T_{k-1} interlace strictly:

$$\lambda_1^{(k)} < \lambda_1^{(k-1)} < \lambda_2^{(k)} < \lambda_2^{(k-1)} < \cdots < \lambda_{k-2}^{(k-1)} < \lambda_{k-1}^{(k)} < \lambda_{k-1}^{(k-1)} < \lambda_k^{(k)}.$$

It will be shown in §3 that, as a function of σ, π_{k-1} vanishes at and only at the $\lambda_i^{(k-1)}, i = 1, \ldots, k-1$, while π_k vanishes at and only at the $\lambda_i^{(k)}, i = 1, \ldots, k$. Since

$$\xi_k = \sqrt{\pi_k^2 + \beta_{k+1}^2} \geq \beta_{k+1}, \qquad k < n,$$

it is apparent that the only entry of R that can ever vanish is $\xi_n = \pi_n$. Consequently,

$$\hat{\beta}_n = \pi_n s_n$$

vanishes at and only at $\sigma = \lambda_i^{(n)}, i = 1, \ldots, n$. To study $\hat{\beta}_n$ for σ in the neighborhood of $\lambda_i^{(n)}$, we consider its derivatives

$$\hat{\beta}_n' = \pi_n' s_n + \pi_n s_n',$$

$$\hat{\beta}_n'' = \pi_n'' s_n + 2\pi_n' s_n' + \pi_n s_n''.$$

Clearly, it is necessary to study the properties of $\pi_k(\sigma)$ to understand $\hat{\beta}_n'$ and $\hat{\beta}_n''$, and this is done in §3.

2.3. Examples of forward instability. We computed the "exact" QR transform by using a very expensive method that works only for singular T as follows. The Schur parameters $\{c_i\}$ are reconstructed from the vector y_n of (2.2), which happens to be an eigenvector of T. These eigenvectors can be computed to high relative accuracy when the eigenvalue is simple, as in our examples.

Example 2.1.

$$(2.19) \quad T = \begin{pmatrix} 6683.3333 & 14899.672 & & & & \\ 14899.672 & 33336.632 & 34.640987 & & & \\ & 34.640987 & 20.028014 & 11.832164 & & \\ & & 11.832164 & 20.001858 & 10.141851 & \\ & & & 10.141851 & 20.002287 & 7.5592896 \\ & & & & 7.5592896 & 20.002859 \end{pmatrix}.$$

Eigenvalues $\lambda_1 = 0$, $\lambda_2 = 10$, $\lambda_3 = 20$, $\lambda_4 = 30$, $\lambda_5 = 40$, $\lambda_6 = 40000$.

(i) Successful deflation with shift $\sigma = 0 = \lambda_1$.

$$\hat{T} = \begin{pmatrix} 39999.925 & 54.726511 & & & & \\ 54.726511 & 33.404823 & 8.3017268 & & & \\ & 8.3017268 & 24.730751 & 8.8065994 & & \\ & & 8.8065994 & 21.646903 & 7.2175779 & \\ & & & 7.2175779 & 20.292461 & -1.113d\text{–}14 \\ & & & & -1.113d\text{–}14 & 9.520d\text{–}13 \end{pmatrix}.$$

The computed version of $\hat{T}$ was indistinguishable from the exact to eight decimals, except that the nonzero entries in the last row are ($-7.943d$–12, $-2.344d$–15).

(ii) Failed deflation with shift $\sigma = \lambda_6 = 40000$.

First, we present the true $\hat{T}$ rounded to eight decimals and then the computed version $\hat{T}^{(1)}$ using double precision on VAX 780.

$$\begin{aligned}
\hat{T} &= \begin{pmatrix}
19.989995 & 14.142133 & & & & \\
14.142133 & 20.003002 & 11.832160 & & & \\
 & 11.832160 & 20.001858 & 10.141851 & & \\
 & & 10.141851 & 20.002287 & 7.5592896 & \\
 & & & 7.5592896 & 20.002859 & -1.608d{-}13 \\
 & & & & -1.608d{-}13 & 40000.000
\end{pmatrix}, \\
\hat{T}^{(1)} &= \begin{pmatrix}
19.989995 & 14.142133 & & & & \\
14.142133 & 20.003002 & 11.832160 & & & \\
 & 11.832160 & 20.001858 & 10.141851 & & \\
 & & 10.141851 & 20.002287 & 7.5593584 & \\
 & & & 7.5593584 & 20.730517 & -170.561 \\
 & & & & -170.561 & 39999.272
\end{pmatrix}.
\end{aligned} \tag{2.20}$$

Insight is gained by looking at the intermediate matrices in the QR sweep. $T^{(4)}$ below shows the premature deflation mentioned in comment 3 in §1 and in (2.18).

Then

$$T^{(1)} = \text{matrix shown in (2.19).}$$

$$T^{(2)} = \begin{pmatrix}
19.989995 & 0.011185926 & 14.142128 & & & \\
0.011185926 & 39999.975 & -31.622748 & & & \\
14.142128 & -31.622748 & 20.028014 & 11.832164 & & \\
 & & 11.832164 & 20.001858 & 10.141851 & \\
 & & & 10.141851 & 20.002287 & 7.5592896 \\
 & & & & 7.5592896 & 20.002859
\end{pmatrix},$$

$$T^{(3)} = \begin{pmatrix}
19.989995 & 14.142133 & & & & \\
14.142133 & 20.003001 & -2.76734d{-}6 & 11.832160 & & \\
 & -2.76734d{-}6 & 40000.000 & 9.35882d{-}3 & & \\
 & 11.832160 & 9.35882d{-}3 & 20.001858 & 10.141851 & \\
 & & & 10.141851 & 20.002287 & 7.5592896 \\
 & & & & 7.5592896 & 20.002859
\end{pmatrix},$$

$$T^{(4)} = \begin{pmatrix}
19.989995 & 14.142133 & & & & \\
14.142133 & 20.003001 & 11.832160 & & & \\
 & 11.832160 & 20.001858 & -8.18031d{-}6 & 10.141851 & \\
 & & -8.18031d{-}6 & 40000.000 & -2.37200d{-}6 & \\
 & & 10.141851 & -2.37200d{-}6 & 20.002287 & 7.5592896 \\
 & & & & 7.5592896 & 20.002859
\end{pmatrix},$$

$$T^{(5)} = \begin{pmatrix}
19.989995 & 14.142133 & & & & \\
14.142133 & 20.003001 & 11.832160 & & & \\
 & 11.832160 & 20.001858 & 10.141851 & & \\
 & & 10.141851 & 20.002287 & 0.032249815 & 7.5592896 \\
 & & & 0.032249815 & 40000.000 & -6.09724d{-}6 \\
 & & & 7.5592896 & -6.09724d{-}6 & 20.002859
\end{pmatrix},$$

$$T^{(6)} = \begin{pmatrix} 19.989995 & 14.142133 & & & & \\ 14.142133 & 20.003002 & 11.832160 & & & \\ & 11.832160 & 20.001858 & 10.141851 & & \\ & & 10.141851 & 20.002287 & 7.5593584 & \\ & & & 7.5593584 & 20.730517 & \mathbf{-170.56153} \\ & & & & \mathbf{-170.56153} & 39999.272 \end{pmatrix}.$$

(iii) A second QR sweep. Finally, we show $\hat{T}^{(2)}$, the QR transform of $\hat{T}^{(1)}$ in (2.20), using the same shift. This shows that, although $\hat{T}^{(2)}$ deflates nicely it is not very close to $\hat{T}$. Thus repeated application of the transform will *not* recover $\hat{T}$ if forward instability occurred.

$$\hat{T}^{(2)} = \begin{pmatrix} 19.979990 & 14.142125 & & & & \\ 14.142125 & 20.006003 & 11.832161 & & & \\ & 11.832161 & 20.003716 & 10.141851 & & \\ & & 10.141851 & 20.004574 & 7.5592897 & \\ & & & 7.5592897 & 20.005717 & 8.425d\text{–}15 \\ & & & & 8.425d\text{–}15 & 40000.000 \end{pmatrix}.$$

Example 2.2 (surprising successful deflation). This example uses a shift that is an exact eigenvalue not only of T but of the odd principal submatrices. Nevertheless, the computed QR transform is perfectly stable:

$$T = \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}.$$

Eigenvalues $\lambda_1 = -2 - \sqrt{3}$, $\lambda_2 = -3$, $\lambda_3 = -2$, $\lambda_4 = -1$, $\lambda_5 = -2 + \sqrt{3}$. With shift $\sigma = -2$,

$$\hat{T} = \begin{pmatrix} -2.0000000 & 1.4142136 & & & \\ 1.4142136 & -2.0000000 & 0.70710678 & & \\ & 0.70710678 & -2.0000000 & 1.2247449 & \\ & & 1.2247449 & -2.0000000 & 0.0000000 \\ & & & 0.0000000 & -2.0000000 \end{pmatrix}.$$

Example 2.3 (W_{21}^-; see [9]).

$$T = W_{21}^- := \text{tridiag}\begin{pmatrix} & 1 & 1 \cdots 1 & 1 & \\ 10 & 9 & \cdots & -9 & -10 \\ & 1 & 1 \cdots 1 & 1 & \end{pmatrix}.$$

Throughout this paper, matrices similar to W_{21}^- but of different sizes are used to best illustrate the properties of the underlying quantities. For simplicity, these matrices are denoted by W_n^- where $n = 2k - 1$, $\alpha_i = k - i$ for $i = 1, 2, \ldots, n$, and $\beta_i = 1$ for $i = 2, \ldots, n$.

This is the generic example of forward instability. W_{21}^- was constructed by Wilkinson for other purposes; see [9]. It has almost uniformly distributed eigenvalues. Rather than exhibit matrices of this size, we plot in Fig. 4 the exact and the computed Schur parameters of the Q matrix

$$Q = \Theta_2^t \Theta_3^t \cdots \Theta_{21}^t,$$

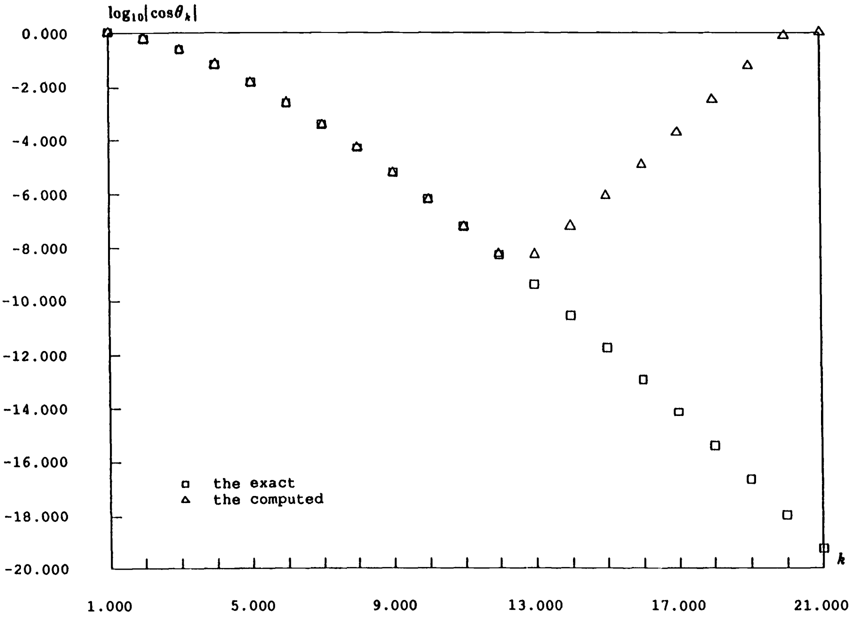


FIG. 4. *Schur parameters defining Q in* QR *transform on W_{21}^{-} (Example* 2.3).

obtained when the largest eigenvalue is the shift. The data in the graph is scaled from x to $\log_{10}|x|$ to show the divergence.

Example 2.4 (forward instability with Wilkinson's shift).

$$T = \text{tridiag}\begin{pmatrix} & 1 & 1\cdots 1 & 1 & \\ 15 & 0 & \cdots & 0 & 15 \\ & 1 & 1\cdots 1 & 1 & \end{pmatrix}.$$

Wilkinson's shift is the eigenvalue of

$$\begin{pmatrix} \alpha_{n-1} & \beta_n \\ \beta_n & \alpha_n \end{pmatrix}$$

that is closer to α_n.

This is another generic example of forward instability. The matrix has a double eigenvalue ($\lambda = 15.066666667$) in double precision. Now we run the QR algorithm on T in single precision with Wilkinson's shift strategy. At the first step Wilkinson's shift is 15.0664. After the first QR sweep, the last off-diagonal element is $-1.96302\mathbf{d}-05$. At the second step Wilkinson's shift is 15.0667. *Premature deflation* is observed at the seventh rotation since (see (2.17)) $\pi_7 * s_7 = -3.98401\mathbf{d}-04$ and $c_7 * \beta_8 = 4.43937\mathbf{d}-03$. To demonstrate forward instability, we repeat the second sweep with the same shift in double precision. The resulting $\hat{\alpha}_k$'s and $\hat{\beta}_k$'s from the second sweep in double precision and in single precision are displayed in Figs. 5 and 6, respectively. Again the data in the graphs is scaled from x to $\log_{10}|x|$ to show the divergence.

3. Properties of π_k. The purpose of this section is to establish properties of π_k that are needed in §4, and consequently §3 may be skipped without loss of continuity.

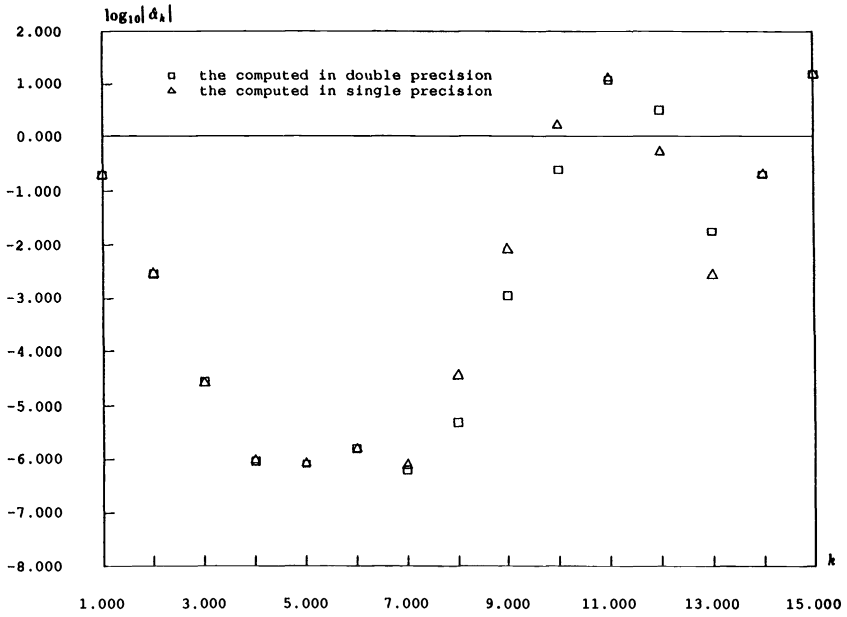


FIG. 5. *Resulting $\hat{\alpha}_k$'s from* QR *sweeps in double precision and in single precision* (*Example* 2.4).

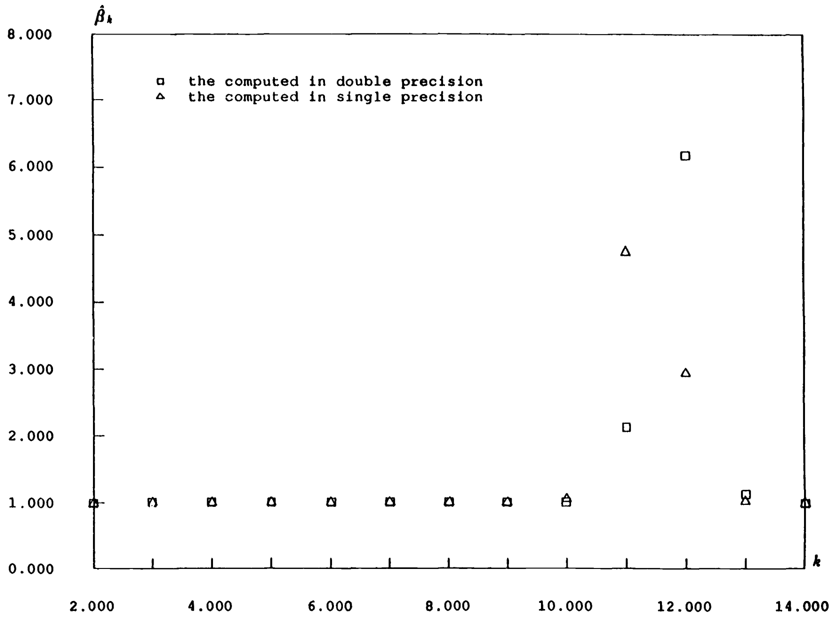


FIG. 6. *Resulting $\hat{\beta}_k$'s from* QR *sweeps in double precision and in single precision* (*Example* 2.4).

3.1. Derivatives of π_k. We discuss the smoothness of $\pi_k(\sigma)$, for $\sigma \in \mathbf{R}$.

LEMMA 3.1. *If T is unreduced (i.e., $\beta_i > 0$, $i = 2, \ldots, n$), then π_k^2 is a rational function with k real double zeros and $2(k-1)$ simple poles that do not lie on the real axis. Thus π_k, c_k, and s_k are real analytic on $\mathbf{R}$. Moreover,*

$$\pi_k(\sigma) = 0 \quad \textit{if and only if } \sigma = \lambda_i^{(k)}, \qquad 1 \le i \le k, 1 \le k \le n. \tag{3.1}$$

The vector y_k of (2.2) is an eigenvector of T_k when $\sigma = \lambda_i^{(k)}$, and

$$c_k(\lambda_i^{(k)}) = (-1)^{i-1}\omega_{ik}, \quad i = 1, 2, \ldots, k, \quad k = 1, 2, \ldots, n. \tag{3.2}$$

Proof. Take the leading $k \times k$ submatrix on each side of (2.1) to find

$$\Theta_k \cdots \Theta_2(T_k - \sigma I_k) = R_k$$

and

$$\det(T_k - \sigma I_k) = \xi_1 \cdots \xi_{k-1}\pi_k. \tag{3.3}$$

By (2.8), $\xi_k \ge \beta_{k+1}$ for all σ and $i < n$, and (3.1) follows directly from (3.3) and the assumption that T is unreduced.

Next consider the submatrix on each side of (2.1) in the first k rows and the first $k-1$ columns:

$$\Theta_k \cdots \Theta_2 \begin{bmatrix} T_{k-1} - \sigma I_{k-1} \\ e_{k-1}^t \beta_k \end{bmatrix} = \begin{bmatrix} \tilde{R}_{k-1} \\ o^t \end{bmatrix}.$$

Since $\Theta_k \cdots \Theta_2$ is orthogonal,

$$(T_{k-1} - \sigma I_{k-1})^2 + e_{k-1}e_{k-1}^t\beta_k^2 = \tilde{R}_{k-1}^t \tilde{R}_{k-1},$$

and, on taking determinants,

$$\det[(T_{k-1} - \sigma I_{k-1})^2 + e_{k-1}e_{k-1}^t\beta_k^2] = (\xi_1 \cdots \xi_{k-1})^2. \tag{3.4}$$

Divide $(3.3)^2$ by (3.4) to find

$$\pi_k^2 = \det{}^2[T_k - \sigma I_k]/\det[(T_{k-1} - \sigma I_{k-1})^2 + \beta_k^2 e_{k-1}e_{k-1}^t].$$

Note that with $\beta_k \ne 0$ the denominator can never vanish for real σ and consideration of the diagonal shows that it is a polynomial of degree $2(k-1)$. The numerator is the square of a polynomial of degree k and thus π_k^2 is a rational function in σ, as claimed.

Note that by (2.9), (2.10), and (2.11), ξ_{k-1}^2, c_k^2, and s_k^2 are rational since π_k^2 is rational and none have poles on the real axis.

Use (2.7) and (3.1) to find that y_k is an eigenvector of T_k when $\sigma = \lambda_i^{(k)}$. Then by definition of y_k in (2.2), find that $\omega_{ik} = |c_k(\lambda_i^{(k)})|$. Use (2.10) to see that c_k has the same sign as π_{k-1} does and $\text{sign}[c_k(\lambda_i^{(k)})] = \text{sign}[\pi_{k-1}(\lambda_i^{(k)})] = (-1)^{i-1}$, since from (3.3), $\pi_{k-1} > 0$ for $\sigma < \lambda_1^{(k-1)}$. That proves (3.2). □

Applying Lemma 3.1, a fundamental relation between π_k, c_k and their derivatives with respect to σ is obtained. Let $f'(\sigma) = df(\sigma)/d\sigma$.

LEMMA 3.2. *For all real σ and $1 \le k \le n$,*

$$\pi_k c_k' - c_k \pi_k' = 1. \tag{3.5}$$

Proof. Differentiate (2.7):

$$-y_k + (T_k - \sigma I_k) y_k' = \pi_k' e_k^{(k)}.$$

Multiply by $(y_k)^t$ and recall the definition of y_k in (2.2):

$$-(y_k)^t y_k + (y_k)^t (T_k - \sigma I_k) y_k' = c_k \pi_k'.$$

Equation (3.5) follows since $(y_k)^t y_k = 1$ by (2.2) and $(y_k)^t (T_k - \sigma I_k) = \pi_k (e_k^{(k)})^t$ by (2.7). □

COROLLARY 1 OF LEMMA 3.2. *For $2 \le k \le n$,*

$$\pi_k'(\lambda_i^{(k)}) = -\frac{1}{c_k(\lambda_i^{(k)})} = \frac{(-1)^i}{\omega_{ik}}, \qquad 1 \le i \le k. \tag{3.6}$$

Proof. Since $\pi_k(\lambda_i^{(k)}) = 0$ by (3.1), relation (3.5) at $\lambda_i^{(k)}$ becomes

$$c_k(\lambda_i^{(k)}) \pi_k'(\lambda_i^{(k)}) = -1.$$

The result follows from (3.2). □

COROLLARY 2 OF LEMMA 3.2.

$$\pi_k''(\lambda_i^{(k)}) = 0, \qquad 1 \le i \le k. \tag{3.7}$$

Proof. Differentiate (3.5): $\pi_k c_k'' - c_k \pi_k'' = 0$. Set $\sigma = \lambda_i^{(k)}$, then $\pi_k = 0$ by (3.1) and $c_k = \pm\omega_{ik} \ne 0$ by (3.2). □

We now turn to the behaviour of π_k for all other values of σ.

LEMMA 3.3. *For T unreduced and $k \le n$,*

$$\pi_k \pi_k'' < 0 \quad \textit{for } \sigma \ne \lambda_i^{(k)}. \tag{3.8}$$

Proof. Premultiply both sides of (2.7) by $(T_k - \sigma I_k)^{-1} \pi_k^{-1}$ to find

$$\frac{y_k}{\pi_k} = (T_k - \sigma I_k)^{-1} e_k^{(k)} \quad \text{for } \sigma \ne \lambda_i^{(k)}. \tag{3.9}$$

A consequence of the spectral factorization is

$$(e_k^{(k)})^t (T_k - \sigma I_k)^{-p} e_k^{(k)} = \sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^p} \quad \text{for } \sigma \ne \lambda_i^{(k)}. \tag{3.10}$$

Since $y_k^t y_k = 1$ by (2.2), (3.9) yields

$$\frac{1}{\pi_k^2} = \left(e_k^{(k)}\right)^t (T_k - \sigma I_k)^{-2} e_k^{(k)}. \tag{3.11}$$

Combine (3.10) and (3.11) to find

$$\frac{1}{\pi_k^2} = \sum_{i=1}^{k} \left(\frac{\omega_{i,k}}{\lambda_i^{(k)} - \sigma} \right)^2 \quad \text{for } \sigma \neq \lambda_i^{(k)}. \tag{3.12}$$

Differentiate both sides of (3.12):

$$-\frac{1}{\pi_k^3}\pi_k' = \sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^3} \quad \text{for } \sigma \neq \lambda_i^{(k)}. \tag{3.13}$$

Differentiate the right-hand side of (3.13):

$$3\sum_{i=1}^{k} \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4}.$$

Differentiate the left-hand side of (3.13):

$$-\frac{1}{\pi_k^3}\pi_k'' + \frac{3}{\pi_k^4}(\pi_k')^2.$$

Therefore, for $\sigma \neq \lambda_i^{(k)}$ with $\sum$ denoting $\sum_{i=1}^{k}$,

$$\begin{aligned}
-\frac{1}{\pi_k^5}\pi_k'' &= \frac{3}{\pi_k^2}\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - \frac{3}{\pi_k^6}(\pi_k')^2 \\
&= 3\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^2} \sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - 3\left(\frac{1}{\pi_k^3}\pi_k'\right)^2 \quad \text{by (3.12)} \\
&= 3\left[\sum \left(\frac{\omega_{i,k}}{\lambda_i^{(k)} - \sigma}\right)^2 \sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - \left(\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^3}\right)^2\right] \quad \text{by (3.13)} \\
&\geq 0 \quad \text{by the Cauchy–Schwarz inequality.}
\end{aligned}$$

Moreover, equality holds if and only if the following two vectors,

$$z_1 = \left(\frac{\omega_{1,k}}{\lambda_1^{(k)} - \sigma}, \ldots, \frac{\omega_{k,k}}{\lambda_k^{(k)} - \sigma} \right)^t,$$

$$z_2 = \left(\frac{\omega_{1,k}}{(\lambda_1^{(k)} - \sigma)^2}, \ldots, \frac{\omega_{k,k}}{(\lambda_k^{(k)} - \sigma)^2} \right)^t,$$

are proportional. That is,

$$\lambda_1^{(k)} - \sigma = \cdots = \lambda_k^{(k)} - \sigma.$$

This is impossible, since T is unreduced. Therefore,

$$-\frac{1}{\pi_k^5}\pi_k''(\sigma) > 0 \quad \text{for } \sigma \neq \lambda_i^{(k)}.$$

Recall from (3.1) that $\pi_k(\sigma) \neq 0$ for $\sigma \neq \lambda_i^{(k)}$. Therefore,

$$-\pi_k\pi_k'' > 0 \quad \text{for } \sigma \neq \lambda_i^{(k)}. \qquad \square$$

COROLLARY OF LEMMA 3.3.

$$\pi_k'' \neq 0 \quad \textit{for } \sigma \neq \lambda_i^{(k)}. \tag{3.14}$$

Lemma 3.3 shows that the algebraic function $\pi_k(\sigma)$ is like the characteristic polynomial of T_k in that it vanishes at the eigenvalues of T_k. Moreover, it is alternatingly concave upward and downward in the intervals bounded by the eigenvalues of T_k. That shows that π_k' attains its extreme values at the $\lambda_i^{(k)}, i = 1, \dots, k$.

LEMMA 3.4. *It holds that*

$$\lim_{\sigma \to \lambda_i^{(k)}} \left(-\frac{\pi_k''}{3\pi_k} \right) = \frac{1}{\omega_{ik}^2} \sum{}' \frac{\omega_{jk}^2}{(\lambda_i^{(k)} - \lambda_j^{(k)})^2}, \tag{3.15}$$

where $\sum'$ *indicates that the term* $j = i$ *is omitted as* $j = 1, \dots, k$.

Proof. Rearrange $-\pi_k''/\pi_k^5$ and use (3.12) to get, with $\sum$ denoting $\sum_{i=1}^k$,

$$\begin{aligned} -\frac{1}{3}\frac{\pi_k''}{\pi_k} &= \pi_k^4 \left[\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^2} \sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - \left(\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^3} \right)^2 \right] \\ &= \frac{\left[\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^2} \sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^4} - \left(\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^3} \right)^2 \right]}{\left(\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^2} \right)^2}. \end{aligned} \tag{3.16}$$

To analyze (3.16) in a neighbourhood of $\lambda_i^{(k)}$, it is convenient to abbreviate the terms. Let

$$\sum \frac{\omega_{i,k}^2}{(\lambda_i^{(k)} - \sigma)^l} = \frac{\omega^2}{\delta^l} + m_l,$$

where $\omega = \omega_{ik}$ and $\delta = \lambda_i^{(k)} - \sigma$. Then (3.16) becomes

$$\frac{(m_2 + \frac{\omega^2}{\delta^2})(m_4 + \frac{\omega^2}{\delta^4}) - (m_3 + \frac{\omega^2}{\delta^3})(m_3 + \frac{\omega^2}{\delta^3})}{(m_2 + \frac{\omega^2}{\delta^2})(m_2 + \frac{\omega^2}{\delta^2})}.$$

Multiply the numerator and denominator by a term δ^4 and simplify the terms to get

$$-\frac{1}{3}\frac{\pi_k''}{\pi_k} = \frac{m_2\omega^2 + O(\delta)}{\omega^4 + O(\delta)}.$$

Then the result follows as $\delta \to 0$. □

COROLLARY OF LEMMA 3.4. *If T is unreduced, then*

$$\pi_k'''(\lambda_i^{(k)}) = (-1)^{i+1} \frac{3}{\omega_{ik}^3} \sum{}' \frac{\omega_{jk}^2}{(\lambda_i^{(k)} - \lambda_j^{(k)})^2}, \tag{3.17}$$

where $\sum'$ *indicates that the term* $j = i$ *is omitted as* $j = 1, \dots, k$.

Proof. Apply L'Hospital's rule:

$$\frac{\pi_k'''(\lambda_i^{(k)})}{\pi_k'(\lambda_i^{(k)})} = \lim_{\sigma\to\lambda_i^{(k)}} \left(\frac{\pi_k''}{\pi_k}\right) = \frac{(-3)}{\omega_{ik}^2}\sum{}' \frac{\omega_{jk}^2}{(\lambda_i^{(k)} - \lambda_j^{(k)})^2}.$$

Using (3.6), the result is obtained. □

The sum in (3.17) plays a role in the local analysis of π_k near each zero $\lambda_i^{(k)}$, $i = 1, \dots, k$.

3.2. The linear model. Near $\lambda_i^{(k)}$, the function $\pi_k(\sigma)$ can be approximated by

$$p_k^{(i)}(\sigma) = \frac{(-1)^i}{\omega_{ik}}(\sigma - \lambda_i^{(k)}). \tag{3.18}$$

We need to know the interval on which this approximation is accurate to within, say, 10 percent.

In this analysis we need the special quantity that appeared in (3.15).

DEFINITION.

$$m_k(i) := 1/\sqrt{\sum{}' \omega_{jk}^2(\lambda_j^{(k)} - \lambda_i^{(k)})^{-2}}, \qquad 1 \le i \le k,$$

where $\sum'$ indicates omission of the term $j = i$ as $j = 1, \dots, k$.

The quantity $m_k(i)$ is almost a mean of the singular values of $T_k - \lambda_i^{(k)} I_k$. Note that

$$\sum{}' \omega_{jk}^2 = 1 - \omega_{ik}^2,$$

and so $m_k(i)^{-2}/(1-\omega_{ik}^2)$ is a weighted average of the $\{(\lambda_j^{(k)} - \lambda_i^{(k)})^{-2}\}_{j\neq i}$. Recall that for positive numbers τ_i, any (weighted) harmonic mean is majorized by the (weighted) arithmetic mean and both lie between the minimum and maximum values, i.e., if $\sum w_i = 1$, $w_i \ge 0$, then

$$\min \tau_i \le \left[\sum w_i \tau_i^{-1}\right]^{-1} \le \sum w_i \tau_i \le \max \tau_i.$$

With $\tau_j = (\lambda_j^{(k)} - \lambda_i^{(k)})^2$, this gives

$$\begin{aligned}\min_{j\neq i}(\lambda_j^{(k)} - \lambda_i^{(k)})^2 &\le \left(\frac{\sum' \omega_{jk}^2(\lambda_j^{(k)} - \lambda_i^{(k)})^{-2}}{\sum' \omega_{jk}^2}\right)^{-1} \\ &\le \frac{\sum' \omega_{jk}^2(\lambda_j^{(k)} - \lambda_i^{(k)})^2}{\sum' \omega_{jk}^2} \le \max_{j\neq i}(\lambda_j^{(k)} - \lambda_i^{(k)})^2.\end{aligned}$$

However,

$$\sum{}' \omega_{jk}^2(\lambda_j^{(k)} - \lambda_i^{(k)})^2 = \sum_{j=1}^{k} \omega_{jk}^2(\lambda_j^{(k)} - \lambda_i^{(k)})^2 = e_k^t(T_k - \lambda_i^{(k)} I)^2 e_k.$$

Thus

$$\min\left\{\lambda_{i+1}^{(k)} - \lambda_i^{(k)}, \lambda_i^{(k)} - \lambda_{i-1}^{(k)}\right\}^2 \le (1 - \omega_{ik}^2) m_k(i)^2$$

and

$$(3.19)\qquad (1-\omega_{ik}^2)m_k(i)^2 \le \frac{e_k^t(T_k-\lambda_i^{(k)}I)^2e_k}{1-\omega_{ik}^2} = \frac{(\alpha_k-\lambda_i^{(k)})^2+\beta_k^2}{1-\omega_{ik}^2}.$$

Now we can compare π_k with the linear function $p_k^{(i)}$.

LEMMA 3.5. *As* $\sigma \longrightarrow \lambda_i^{(k)}$,

$$(3.20)\qquad \frac{\pi_k}{p_k^{(i)}} = 1 - \frac{1}{2}\left(\frac{\sigma-\lambda_i^{(k)}}{\omega_{ik}m_k(i)}\right)^2 + O((\sigma-\lambda_i^{(k)})^3).$$

Proof. Use (3.6), (3.7), and (3.17) to find that the Taylor series of π_k around $\lambda_i^{(k)}$ is

$$\begin{aligned}\pi_k(\sigma) &= 0 + \pi_k'(\lambda_i^{(k)})(\sigma-\lambda_i^{(k)}) + 0 + \frac{1}{6}\pi_k'''(\lambda_i^{(k)})(\sigma-\lambda_i^{(k)})^3 + \cdots\\ &= \frac{(-1)^{k-i}}{\omega_{ik}}(\sigma-\lambda_i^{(k)}) + \frac{1}{6}\frac{(-1)^{k-i-1}3(\sigma-\lambda_i^{(k)})^3}{\omega_{ik}^3 m_k(i)^2} + O((\sigma-\lambda_i^{(k)})^4)\\ &= p_k^{(i)}(\sigma)\left[1-\frac{1}{2}\left(\frac{\sigma-\lambda_i^{(k)}}{\omega_{ik}m_k(i)}\right)^2 + O((\sigma-\lambda_i^{(k)})^3)\right]. \qquad \square\end{aligned}$$

Remark 1. It is not valid to use the linear model for π_k if the quadratic term in (3.20) exceeds 1. Thus, in the generic case when there is no happy cancellation between the quadratic and higher terms, it is *necessary* to require

$$|\sigma-\lambda_i^{(k)}| < \tfrac{1}{2}\omega_{ik}m_k(i)$$

so that

$$\frac{\pi_k}{p_k^{(i)}} > \frac{7}{8} + O((\sigma-\lambda_i^{(k)})^3).$$

In Fig. 7, the graphs of π_k and $p_k^{(i)}$ are shown as well as the boundary points

$$\lambda_i^{(k)} \pm \tfrac{1}{2}\omega_{ik}m_k(i)$$

in a typical case.

Remark 2. Since

$$m_k(i)^{-2} \le (k-1)\max_{j\ne i}\left(\frac{\omega_{jk}}{\lambda_j^{(k)}-\lambda_i^{(k)}}\right)^2,$$

it follows that

$$\frac{|\lambda_i^{(k)}-\sigma|}{\omega_{ik}m_k(i)} \le \frac{\sqrt{k-1}|\lambda_i^{(k)}-\sigma|}{\omega_{ik}\min_{j\ne i}\left(\frac{|\lambda_j^{(k)}-\lambda_i^{(k)}|}{\omega_{jk}}\right)},$$

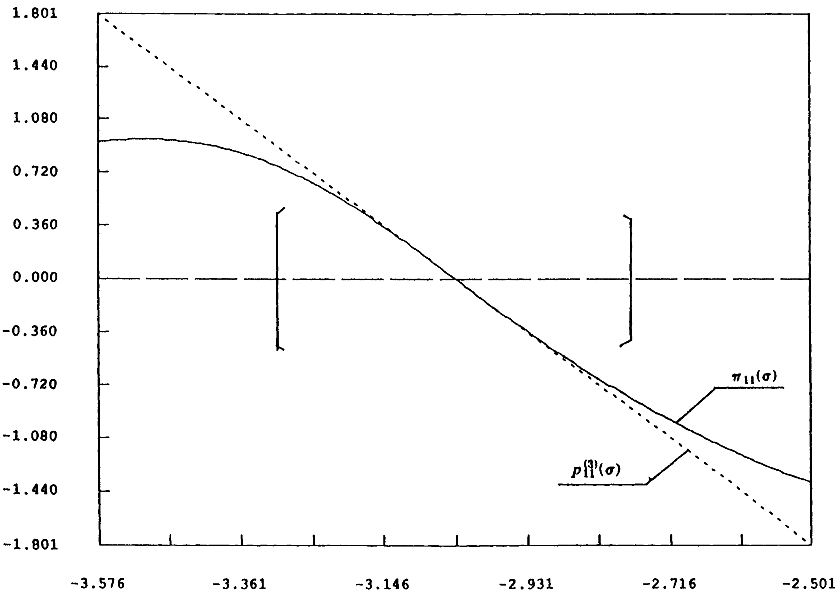


FIG. 7. *π_{11} and the linear model $p_{11}^{(3)}$ on $[\lambda_3^{(11)}-\omega_{3,11}m_{11}(3), \lambda_3^{(11)}+\omega_{3,11}m_{11}(3)]$ with $T = W_{11}^-$ (see* §3.2).

and when the cubic terms are negligible, the constraint

$$|\lambda_i^{(k)} - \sigma| \le \frac{\omega_{ik}}{3\sqrt{k-1}} \min_{j\ne i} \left(\frac{|\lambda_j^{(k)} - \lambda_i^{(k)}|}{\omega_{jk}} \right)$$

is a *sufficient* condition for using the linear model. Although the minimum will not be known, the expression shows how a small value of ω_{jk} tends to neutralize the effect of close $\lambda_j^{(k)}$ in restricting the domain of the linear model.

In the analysis of $\hat{\beta}_k'$ and $\hat{\alpha}_k'$ near $\lambda_i^{(k)}$, the following auxiliary function plays a role:

$$\phi_k := \pi_k\pi_k'/\xi_k^2 = \pi_k\pi_k'/(\pi_{k-1}^2 + \beta_k^2).$$

See Figs. 2 and 3 for the shape of ϕ_k near $\lambda_i^{(k)}$.

LEMMA 3.6. *When $|\sigma - \lambda_i^{(k)}| < \frac{1}{2}\omega_{ik}m_k(i)$ and $k < n$, then*

$$\phi_k/f_k^{(i)} \in (1/2, 1]$$

where

$$f_k^{(i)}(\sigma) := \frac{\sigma - \lambda_i^{(k)}}{\beta_{k+1}^2\omega_{ik}^2 + (\sigma - \lambda_i^{(k)})^2}, \qquad 1 \le i \le k. \tag{3.21}$$

Proof. Abbreviate $\sigma - \lambda_i^{(k)}$ by δ, ω_{ik} by ω, $m_k(i)$ by m, and use Lemma 3.5 to find

$$\begin{aligned}\frac{\phi_k}{f_k^{(i)}} &= \frac{1 - 2(\frac{\delta}{\omega m})^2 + O(\delta^3)}{1 - (\frac{\delta}{\omega m})^2 \frac{\delta^2}{\omega^2\beta^2+\delta^2} + O(\delta^5)} \\ &= 1 - \left(2 - \frac{\delta^2}{\omega^2\beta^2+\delta^2}\right)\left(\frac{\delta}{\omega m}\right)^2 + O(\delta^3).\end{aligned}$$

Thus, when $|\delta| < \frac{1}{2}\omega m$,

$$0 \leq -\frac{\phi_k}{f_k^{(i)}} + 1 + O(\delta^3) < \frac{1}{2}.$$

Equality on the left occurs only when $\delta = 0$. □

COROLLARY.

$$|f_k^{(i)}| \leq \frac{1}{2\beta_{k+1}\omega_{ik}}$$

with equality if and only if $\sigma = \lambda_i^{(k)} \pm \omega_{ik}\beta_{k+1}$.

Remark. If $\beta_{k+1} > m_k(i)$, then $|\phi_k|$ will not attain the bound on $|f_k^{(i)}|$ since $|\phi_k| < |f_k^{(i)}|$ on the interval in Lemma 3.6, and this interval does not contain the maximizing point of $|f_k^{(i)}|$.

3.3. Pointwise bounds. We now give a pointwise bound on π_k' that reveals the role of the distance of σ from the spectrum of T_k and from the spectrum of T_{k-1}. Some preliminary results are restated for easy reference.

(A) $\xi_{k-1} = \sqrt{\pi_{k-1}^2 + \beta_k^2}$ (see definition in (2.9)),

(B) $c_k = \pi_{k-1}/\xi_{k-1}$ (see definition in (2.10)),

(C) $s_k = \beta_k/\xi_{k-1}$ (see definition in (2.11)),

(D) $c_k' = \frac{s_k^2}{\xi_{k-1}}\pi_{k-1}'$ (derivative of (B)),

(E) $\pi_k c_k' - c_k \pi_k' = 1$ (for any real σ (3.5)),

(F) $(T_k - \sigma I_k)y_k = \pi_k e_k^{(k)}$ (for any real σ (2.7)).

DEFINITION. $\mu_k := \mu_k(\sigma) = \min_{1 \leq i \leq k} |\lambda_i^{(k)} - \sigma|, k = 1, 2, \ldots, n.$

$$\text{(G)} \qquad \mu_k^2 < \pi_k^2 \quad (\text{for } \sigma \neq \lambda_i^{(k)}).$$

Proof of (G). $(F)^t(F)$ yields $(y_k)^t(T_k - \sigma I_k)^2 y_k = \pi_k^2$. Then

$$\begin{aligned}\mu_k^2 &= \min_{1\leq i \leq k}(\sigma - \lambda_i^{(k)})^2 \quad \text{(by definition)} \\ &= \lambda_{\min}((T_k - \sigma I_k)^2) \\ &< (y_k)^t(T_k - \sigma I_k)^2 y_k \quad \text{since } y_k \text{ cannot be an eigenvector since } \sigma \neq \lambda_i^{(k)} \\ &= \pi_k^2. \quad \square\end{aligned}$$

The next result shows that $|\pi_k'|$ can only be huge when $\mu_k(\sigma)$ and $\mu_{k-1}(\sigma)$ are both tiny.

LEMMA 3.7. *For any real* σ,

$$\mu_k(\sigma)|\pi_k'| = |\pi_k|, \qquad \sigma = \lambda_i^{(k)}, \tag{3.22}$$

$$\mu_k(\sigma)|\pi_k'| < |\pi_k|, \qquad \sigma \neq \lambda_i^{(k)}, \tag{3.23}$$

$$\mu_{k-1}(\sigma)|\pi_k'| < s_k^2|\pi_k| + \sqrt{\mu_{k-1}^2(\sigma) + \beta_k^2}. \tag{3.24}$$

Proof of (3.22), (3.23). When $\sigma = \lambda_i^{(k)}$, then $\pi_k = 0$ by (3.1), $\mu_k = 0$ by definition, and (3.22) holds. Now suppose that $\sigma \neq \lambda_i^{(k)}$. Premultiply (F) by $(T_k - \sigma I_k)^{-1}\pi_k{}^{-1}$:

$$\frac{y_k}{\pi_k} = (T_k - \sigma I_k)^{-1} e_k^{(k)}. \tag{3.25}$$

Differentiate (3.25):

$$\frac{y_k'\pi_k - y_k\pi_k'}{\pi_k^2} = (T_k - \sigma I_k)^{-2} e_k^{(k)}. \tag{3.26}$$

Differentiate $y_k^t y_k = 1$ to obtain $y_k^t y_k' = 0$. Premultiply (3.26) by y_k^t, getting

$$\begin{aligned} -\frac{1}{\pi_k^2}\pi_k' &= (y_k)^t (T_k - \sigma I_k)^{-2} e_k^{(k)} \quad \text{since } y_k^t y_k' = 0 \\ &= (y_k)^t (T_k - \sigma I_k)^{-1} y_k/\pi_k \quad \text{by (3.25)}. \end{aligned}$$

Thus

$$-\pi_k' = (y_k)^t (T_k - \sigma I_k)^{-1} y_k \pi_k$$

and

$$|\pi_k'| = |(y_k)^t (T_k - \sigma I_k)^{-1} y_k||\pi_k|. \tag{3.27}$$

Since $\sigma \neq \lambda_i^{(k)}$, y_k will not be an eigenvector. Therefore,

$$\begin{aligned} |(y_k)^t (T_k - \sigma I_k)^{-1} y_k| &< \max_{v^t v = 1} |v^t (T_k - \sigma I_k)^{-1} v| \\ &= \|(T_k - \sigma I_k)^{-1}\| \\ &= \frac{1}{\min_{1 \le i \le k} |\lambda_i^{(k)} - \sigma|} = \frac{1}{\mu_k}. \end{aligned}$$

Put this inequality into (3.27) and multiply by μ_k to obtain (3.23). □

Proof of (3.24). When $\sigma = \lambda_i^{(k-1)}$, then $\mu_{k-1} = 0$, and (3.24) is immediate since $\beta_k \neq 0$ and $s_k^2|\pi_k| \neq 0$ by (3.1).

Now, suppose that $\sigma \neq \lambda_i^{(k-1)}$. Then,

$$\begin{aligned} |\pi_k' c_k + 1| &= |c_k' \pi_k| \quad \text{by (E)} \\ &= \frac{s_k^2}{\xi_{k-1}} |\pi_{k-1}' \pi_k| \quad \text{by (D)} \\ &< \frac{s_k^2}{\xi_{k-1}} \frac{|\pi_{k-1}|}{\mu_k} - 1|\pi_k| \quad \text{by (3.23)} \\ & \frac{{}^{|}c_k \pi_k|}{-1} \quad \text{by (B)}. \end{aligned}$$

Hence,

$$|\pi_k' c_k| < \frac{s_k^2|c_k\pi_k|}{\mu_{k-1}} + 1.$$

Since $\sigma \neq \lambda_i^{(k-1)}$, $c_k \neq 0$ by (3.1). Then the above inequality can be rearranged as

$$\begin{aligned}
\mu_{k-1}|\pi_k'| &< s_k^2|\pi_k| + \frac{\mu_{k-1}}{|c_k|} \\
&= s_k^2|\pi_k| + \mu_{k-1}\frac{\xi_{k-1}}{|\pi_{k-1}|} \quad \text{by (B)} \\
&= s_k^2|\pi_k| + \mu_{k-1}\frac{\sqrt{\pi_{k-1}^2 + \beta_k^2}}{|\pi_{k-1}|} \quad \text{by (A)} \\
&= s_k^2|\pi_k| + \sqrt{\mu_{k-1}^2 + \beta_k^2\frac{\mu_{k-1}^2}{\pi_{k-1}^2}} \\
&< s_k^2|\pi_k| + \sqrt{\mu_{k-1}^2 + \beta_k^2} \quad \text{by (G)}. \qquad \square
\end{aligned}$$

4. Sensitivity of $\hat{T}$. Recall that $\hat{T} = RQ + \sigma I = Q^tTQ$ where $T - \sigma I = QR$. In terms of the intermediate quantities generated in the QR transformation, the entries of $\hat{T}$ are given in (2.14) and (2.15), namely,

$$\begin{aligned}
\hat{\beta}_k &= \xi_k s_k = \beta_k(\xi_k/\xi_{k-1}) \quad \text{for } k < n \quad \text{and} \quad \xi_k^2 = \pi_k^2 + \beta_{k+1}^2, \\
\hat{\alpha}_k &= \alpha_{k+1} - \gamma_{k+1} + \gamma_k \quad \text{for } k < n \quad \text{and} \quad \gamma_k = \pi_k c_k, \\
\hat{\beta}_n &= \pi_n s_n, \qquad \hat{\alpha}_n = \gamma_n + \sigma.
\end{aligned} \tag{4.0}$$

Differentiating with respect to σ reveals that

$$\hat{\beta}_k' = \xi_k' s_k + \xi_k s_k' = \beta_k(\xi_k/\xi_{k-1})', \tag{4.1}$$

$$\hat{\alpha}_k' = \gamma_k' - \gamma_{k+1}', \tag{4.2}$$

$$|\hat{\beta}_n'| = |\pi_n' s_n + \pi_n s_n'|, \tag{4.3}$$

$$\hat{\alpha}_n' = \gamma_n' + 1. \tag{4.4}$$

Here is the plan of this section. Our first goal is to show that these derivatives cannot be huge unless σ is close to an eigenvalue of some T_k. Next we show that, in fact, to have "huge" derivatives, σ must be close to the spectrum of two successive T_k's. The proofs repeatedly use two results, (2.7) and (3.23), from the previous sections:

$$c_k'\pi_k - c_k\pi_k' = 1 \quad \text{for all } \sigma \quad \text{and} \quad k = 2, \ldots, n, \tag{4.5}$$

and

$$|\pi_k'| < |\pi_k|/\mu_k \quad \text{for } \sigma \notin \text{spectrum } (T_k). \tag{4.6}$$

Here

$$\mu_k = \min_{1\le i\le k} |\sigma - \lambda_i^{(k)}| = \operatorname{dist}(\sigma, \operatorname{spectrum}(T_k)). \tag{4.7}$$

Finally, we give realistic bounds on all these derivatives.

The first goal is met by the following result.

THEOREM 4.1. *With $\mu_k(\sigma)$ defined in* (4.7) *and $k < n$,*

$$\text{(i)}\quad |\hat\beta_k'| < \hat\beta_k\left(\frac{c_{k+1}^2}{\mu_k} + \frac{c_k^2}{\mu_{k-1}}\right), \quad \sigma \notin \operatorname{spectrum}(T_k) \cup \operatorname{spectrum}(T_{k-1}), \tag{4.8}$$

$$\text{(ii)}\quad |\hat\alpha_k'| < 2\left(\frac{|\gamma_k|}{\mu_k} + \frac{|\gamma_{k-1}|}{\mu_{k+1}}\right), \quad \sigma \notin \operatorname{spectrum}(T_{k+1}) \cup \operatorname{spectrum}(T_k), \tag{4.9}$$

$$\text{(iii)}\quad |\hat\beta_n'| < \hat\beta_n\left(\frac{1}{\mu_n} + \frac{c_n^2}{\mu_{n-1}}\right), \quad \sigma \notin \operatorname{spectrum}(T_n) \cup \operatorname{spectrum}(T_{n-1}), \tag{4.10}$$

$$\text{(iv)}\quad |\hat\alpha_n'| < 2\left(\frac{|\gamma_n|}{\mu_n} + 1\right), \quad \sigma \notin \operatorname{spectrum}(T_n). \tag{4.11}$$

Proof.

$$\text{(i)}\qquad \frac{\xi_k}{\xi_{k-1}} = \sqrt{\frac{\pi_k^2 + \beta_{k+1}^2}{\pi_{k-1}^2 + \beta_k^2}}.$$

Applying the quotient rule and simplifying the result yields

$$\left(\frac{\xi_k}{\xi_{k-1}}\right)' = \left(\frac{\xi_k}{\xi_{k-1}}\right)\left(\frac{\pi_k\pi_k'}{\xi_k^2} - \frac{\pi_{k-1}\pi_{k-1}'}{\xi_{k-1}^2}\right).$$

Use (4.6) to get

$$\left|\left(\frac{\xi_k}{\xi_{k-1}}\right)'\right| < \left(\frac{\xi_k}{\xi_{k-1}}\right)\left(\frac{\pi_k^2}{\xi_k^2\mu_k} + \frac{\pi_{k-1}^2}{\xi_{k-1}^2\mu_{k-1}}\right)$$
$$= \frac{\xi_k}{\xi_{k-1}}\left(\frac{c_{k+1}^2}{\mu_k} + \frac{c_k^2}{\mu_{k-1}}\right).$$

Multiply each side by β_k and use (4.1) to obtain (4.8).

(ii) From (4.2) and (4.0),

$$\hat\alpha_k' = \gamma_k' - \gamma_{k+1}' = c_k'\pi_k + c_k\pi_k' - c_{k+1}'\pi_{k+1} - c_{k+1}\pi_{k+1}'.$$

Now use (4.5) to conclude $\hat\alpha_k' = 1 + 2c_k\pi_k' - (1 + 2c_{k+1}\pi_{k+1}')$, and use (4.6) to find

$$|\hat\alpha_k'| < 2\left(\frac{|c_k\pi_k|}{\mu_k} + \frac{|c_{k+1}\pi_{k+1}|}{\mu_{k+1}}\right) \quad \text{for } \sigma \notin \operatorname{spectrum}(T_k) \cup \operatorname{spectrum}(T_{k+1}).$$

$$\text{(iii)}\qquad s_n' = (\beta_n/\xi_{n-1})' = -\frac{\beta_n\pi_{n-1}\pi_{n-1}'}{\xi_{n-1}^3} = -s_nc_n\frac{\pi_{n-1}'}{\xi_{n-1}}. \tag{4.12}$$

$$\begin{aligned}|\hat{\beta}_n'| &= |\pi_n' s_n - \pi_n s_n c_n \pi_{n-1}'/\xi_{n-1}| \\ &< |\pi_n| s_n \left(\frac{1}{\mu_n} + \frac{|c_n \pi_{n-1}|}{\xi_{n-1}\mu_{n-1}}\right) \quad \text{for } \sigma \notin \text{spectrum}(T_n) \cup \text{spectrum}(T_{n-1}), \\ &= \hat{\beta}_n \left(\frac{1}{\mu_n} + \frac{c_n^2}{\mu_{n-1}}\right).\end{aligned}$$

$$\text{(iv)} \qquad \hat{\alpha}_n' = \gamma_n' + 1 = \begin{cases} 2\pi_n c_n' \\ 2(\pi_n' c_n + 1) \end{cases} \qquad \text{since } c_n' \pi_n = 1 + c_n \pi_n'.$$

Then

$$|\hat{\alpha}_n'| < \begin{cases} 2|\gamma_n|/\mu_{n-1} \\ 2\left(\dfrac{|\pi_n c_n|}{\mu_n} + 1\right) \end{cases} \qquad \text{for } \sigma \notin \text{spectrum}(T_n) \cup \text{spectrum}(T_{n-1}). \qquad \Box$$

It may be verified that the $\gamma_i, i = 1, \ldots, n$ are intermediate quantities that appear on the diagonal when $T - \sigma I$ is being transformed into $\hat{T}$ by a sequence of plane rotations. Hence $|\gamma_i| < \|T - \sigma I\|$ for $i = 1, \ldots, n$ and all σ. Thus huge values for the derivative of $\hat{T}$ can come only from tiny values of some μ_k.

The next step is to show that the derivatives at the points excluded in the previous theorem are of modest size *unless* the excluded points, in each case, are very close to each other.

In order to simplify notation we shall label eigenvalues of the submatrices T_k so that the closest eigenvalue of T_{k-1} to $\lambda_i^{(k)}$ is $\lambda_{i_-}^{(k-1)}$ and of T_{k+1} is $\lambda_{i_+}^{(k+1)}$. Formally,

$$(*) \qquad \mu_{k-1}(\lambda_i^{(k)}) = |\lambda_{i_-}^{(k-1)} - \lambda_i^{(k)}|, \qquad \mu_{k+1}(\lambda_i^{(k)}) = |\lambda_{i_+}^{(k+1)} - \lambda_i^{(k)}|.$$

In all cases, i_- is either i or $i - 1$ and i_+ is either i or $i + 1$.

THEOREM 4.2. *With the notation above, for $i = 1, 2, \ldots, k$,*

$$\begin{aligned}
&\text{(i)} && |\hat{\beta}_k'(\lambda_{i_-}^{(k-1)})| = \hat{\beta}_k \left|\tfrac{\pi_k \pi_k'}{\xi_k^2}\right|_{\lambda_{i_-}^{(k-1)}} < \hat{\beta}_k \tfrac{c_{k+1}^2}{|\lambda_{i_-}^{(k-1)} - \lambda_i^{(k)}|}, \\
&\text{(ii)} && |\hat{\beta}_k'(\lambda_i^{(k)})| = \left| - \hat{\beta}_k \tfrac{\pi_{k-1}\pi_{k-1}'}{\xi_{k-1}^2}\right|_{\lambda_i^k} < \hat{\beta}_k \tfrac{c_k^2}{|\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}|}, \\
&\text{(iii)} && |\hat{\alpha}_k'(\lambda_i^{(k)})| = |2c_k \pi_k'|_{\lambda_i^{(k)}} = 2, \\
&\text{(iv)} && |\hat{\alpha}_k'(\lambda_{i_+}^{(k+1)})| = |2(c_k \pi_k'|_{\lambda_{i_+}^{(k+1)}} + 1)| < 2\left(1 + \tfrac{|\gamma_k|}{|\lambda_{i_+}^{(k+1)} - \lambda_i^{(k)}|}\right), \\
&\text{(v)} && |\hat{\beta}_n'(\lambda_i^{(n)})| = |s_n \pi_n'|_{\lambda_i^{(n)}} < \sqrt{\tfrac{\Delta^2 + \beta_n^2}{\Delta}}; \qquad \Delta = |\lambda_i^{(n)} - \lambda_{i_-}^{(n-1)}|, \\
&\text{(vi)} && |\hat{\beta}_n'(\lambda_{i_-}^{(n-1)})| = |\pi_n' s_n|_{\lambda_{i_-}^{(n-1)}} < \tfrac{\hat{\beta}_n}{\Delta}, \\
&\text{(vii)} && \hat{\alpha}_n'(\lambda_i^{(n)}) = 0, \\
&\text{(viii)} && |\hat{\alpha}_n'(\lambda_i^{(n-1)})| = 2.
\end{aligned}$$

Proof. (i) and (ii) Recall that

$$\hat{\beta}_k' = \beta_k \left(\frac{\pi_k \pi_k'}{\xi_k^2} - \frac{\pi_{k-1}\pi_{k-1}'}{\xi_{k-1}^2}\right)$$

and note that one of the two terms vanishes at the evaluation points and the other may be bounded in the same way as in the previous theorem.

(iii) Recall that $\hat{\alpha}_k' = 2(c_k\pi_k' - c_{k+1}\pi_{k+1}')$ and note that one term evaluates to -1 at the evaluation point since, for example, $c_k\pi_k' - c_k'\pi_k = -1$ and π_k vanishes at $\lambda_i^{(k)}$. Also, $c_{k+1} = 0$ at $\lambda_i^{(k)}$.

(iv) Recall that $\hat{\alpha}_k' = 2+2c_k\pi_k' - 2c_{k+1}'\pi_{k+1}$, and note that π_{k+1} vanishes at $\lambda_{i_+}^{(k+1)}$ and that the other may be bounded in the same way as in the previous theorem.

(v) and (vi) The second term $\pi_n s_n'$ vanishes at both evaluation points. To bound π_n' at $\lambda_k^{(n)}$ we invoke the inequality in (3.24) from the previous section. To bound π_n' at $\lambda_{i_-}^{(n-1)}$ we use the proof of Theorem 4.1 as before.

(vii) and (viii) $\hat{\alpha}_n' = 2\pi_n c_n' = 2(1 + \pi_n' c_n)$ and $\pi_n(\lambda_i^{(n)}) = 0$, $c_n(\lambda_{i_-}^{(n-1)}) = 0$. □

Preamble to Theorem 4.3. Theorem 4.1 shows that it is necessary for σ to be close to some $\lambda_i^{(k)}$, $i = 1, 2, \ldots, k$ in order for $\hat{\beta}_k$ and $\hat{\alpha}_k$ to have large derivatives. Theorem 4.2 shows that $\lambda_i^{(k)}$ must be very close to $\lambda_{i_-}^{(k-1)}$ in order for the derivatives to be large at these points. It remains to show that this last condition is also sufficient for extreme sensitivity of $\hat{T}$ to changes in σ. Indeed $|\hat{\beta}_k'|$, for $k < n$, may be modest at both $\lambda_i^{(k)}$ and $\lambda_{i_-}^{(k-1)}$ even when they are extremely close, and it is only at other values of σ in the neighborhood of $\lambda_i^{(k)}$ that the large values occur. Theorem 4.3 establishes this fact.

What permits this analysis is that in an interval of width $m_k(i)\omega_{ik}$ centered on $\lambda_i^{(k)}$ (see the linear model discussion in §3 for the definition of $m_k(i)$), both $\hat{\alpha}_k'$ and $\hat{\beta}_k'$ may be approximated closely by much simpler functions of σ whose maxima can be estimated. Recall that ω_{ik} is the magnitude of the last entry in $\lambda_i^{(k)}$'s normalized eigenvector. One difficulty in the analysis is to pin down $\lambda_{i_-}^{(k-1)}$, the closest eigenvalue of T_{k-1}, which must lie within the interval that we just associated with $\lambda_i^{(k)}$. Theorem 4.4, quoted at the end of this section, shows that

$$|\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}| = O(\omega_{ik}^2),$$

except in special circumstances when it is only $O(\omega_{ik})$, the width of our interval.

We digress to explain the situation. Lemma 2.1 shows that

$$\|Ty_k^{(n)} - y_k^{(n)}\lambda_i^{(k)}\| = \sqrt{\pi_k^2 + \beta_{k+1}^2 c_k^2} = \beta_{k+1}\omega_{ik},$$

since $\pi_k(\lambda_i^{(k)}) = 0$ and $|c_k(\lambda_i^{(k)})| = \omega_{ik}$. Standard results on symmetric matrices (see [4]) let us conclude that there is an eigenvalue λ of T satisfying $|\lambda - \lambda_i^{(k)}| \leq \beta_{k+1}\omega_{ik}$. Moreover, the same can be said for some eigenvalue λ for each T_j with $j > k$. When $\omega_{ik} << 1$, we say that $\lambda_i^{(k)}$ has "stabilized" as an approximate eigenvalue of T. In many cases, there will only be one eigenvalue of T_{k-1} close to $\lambda_i^{(k)}$ and, in that case, their separation is $O(\omega_{ik}^2)$. Theorem 4.4 gives

$$|\Delta| := |\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}| < \omega_{ik}^2 \begin{cases} \lambda_k^{(k)} - \lambda_1^{(k)}, & i = 1, k, \\ \dfrac{(\lambda_k^{(k)} - \lambda_i^{(k)})(\lambda_i^{(k)} - \lambda_1^{(k)})}{|\lambda_i^{(k)} - \lambda_{i_*}^{(k-1)}|}, & 1 < i < k. \end{cases}$$

Here i_* is the second closest eigenvalue and $i_- + i_* = 2i - 1$. Thus, in all cases when $i \neq 1, k$,

$$|\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}| \leq \sqrt{|\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}||\lambda_i^{(k)} - \lambda_{i_*}^{(k-1)}|}$$
$$< \omega_{ik}((\lambda_k^{(k)} - \lambda_i^{(k)})(\lambda_i^{(k)} - \lambda_1^{(k)}))^{1/2} = O(\omega_{ik}).$$

In the special cases when $|\lambda_i^{(k)} - \lambda_{i_*}^{(k-1)}|$ is comparable to Δ, our simple model for π_k is still applicable, but the maxima of the derivatives of $\hat{\alpha}_k$ and $\hat{\beta}_k$ are not so large, in general, and are complicated to express. In fact, the large derivatives will occur for a smaller value of k when the first stabilization occurs. That is why we make the assumption $\Delta = O(\omega_{ik}^2)$ in Theorem 4.3. A simple version of Theorem 4.3 is given next. The distribution of max and min was a surprise to us.

Near $\sigma = \lambda_i^{(k)}$,

$$\text{(i)} \quad \max_\sigma |\hat{\beta}_n'| = \frac{1}{\omega_{\text{in}}}, \qquad k = n,$$

$$\text{(ii)} \quad \max_\sigma |\hat{\beta}_k'| = O\left(\frac{1}{\min(\omega_{ik}, \omega_{i_-,k-1})}\right), \qquad k < n,$$

$$\text{(iii)} \quad \max_\sigma |\hat{\alpha}_n'| = O\left(\frac{1}{\max(\omega_{ik}, \omega_{i_-,k-1})}\right), \qquad k = n,$$

$$\text{(iv)} \quad \max_\sigma |\hat{\alpha}_k'| = O\left(\frac{1}{\min\left\{\max(\omega_{ik}, \omega_{i_-,k-1}), \max(\omega_{ik}, \omega_{i_+,k+1})\right\}}\right), \qquad k < n.$$

The function of Theorem 4.3 is to provide the constants hidden by the O. An interesting byproduct of the proof of Theorem 4.3 is that only for $|\hat{\beta}_n'|$ does the maximum occur in the tiny interval $[\lambda_i^{(n)}, \lambda_{i_-}^{(n-1)}]$. We remind the reader that from the preamble to the proof of Lemma 3.5

$$m_k(i) = 1 \Bigg/ \sqrt{\sum_{j \neq i} \omega_{jk}^2 (\lambda_j^{(k)} - \lambda_i^{(k)})^{-2}}.$$

THEOREM 4.3. *Near $\sigma = \lambda_i^{(k)}$, for $2 \leq k \leq n$ and $1 \leq i \leq k$, if*

$$\Delta_i^k := |\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}| < 100(\lambda_k^{(k)} - \lambda_1^{(k)})\omega_{ik}^2 = O(\omega_{ik}^2),$$

then the entries in $\frac{d}{d\sigma}\hat{T}$ achieve the values indicated below.

$$\text{(i)} \quad \max_\sigma |\hat{\beta}_n'| = \frac{1}{\omega_{in}} \left[1 + \frac{1}{8}\left(\frac{\Delta_i^n}{\beta_n \omega_{i_-,n-1}}\right)^2 - \frac{3}{8}\left(\frac{\Delta_i^n}{\omega_{in} m_n(i)}\right)^2 + O(\Delta^3)\right]$$

and occurs near $(\lambda_i^{(n)} + \lambda_{i_-}^{(n-1)})/2$.

(ii) *The function $\hat{\beta}_k'$, $k < n$, is the difference of two similarly shaped functions; see Fig. 2. Under mild conditions, stated in the proof, the magnitude of the first term's model, on the appropriate interval, exceeds*

$$\frac{1}{2\omega_{ik}}.$$

Under mild conditions, stated in the proof, the magnitude of the second term's model, on the appropriate interval, exceeds

$$\left(\frac{\beta_{k+1}}{\sqrt{2}\beta_k}\right)\frac{1}{2\omega_{i_-,k-1}}.$$

When one of these terms dominates, then the larger one gives an accurate estimate of $\max|\hat{\beta}_k'|$ *near* $\lambda_i^{(k)}$, *under the stated conditions. In all other cases, the maximum value of the model is a more complicated expression and has a smaller value.*

(iii) *The function* $|\hat{\alpha}_n'|$ *attains its maximum, on the intervals associated with* $\lambda_i^{(n)}$ *and* $\lambda_{i_-}^{(n-1)}$, *at the ends of those intervals. There the model achieves*

$$\begin{cases} \dfrac{2}{\omega_{i_-,n-1}}\dfrac{m_n(i)}{\sqrt{4\beta_n^2+m_n^2(i)}} & \textit{if } m_n(i)\omega_{in} < m_{n-1}(i_-)\omega_{i_-,n-1}, \\ \dfrac{2}{\omega_{in}}\dfrac{m_{n-1}(i_-)}{\sqrt{4\beta_n^2+m_{n-1}^2(i_-)}} & \textit{otherwise.} \end{cases}$$

(iv) *The function* $\hat{\alpha}_k'$, $k < n$ *is the difference of two similarly shaped functions. On the appropriate interval about* $\lambda_i^{(k)}$, $i = 1, \dots, k$, *the first term's model, in magnitude, achieves*

$$\begin{cases} \dfrac{2}{\omega_{i_-,k-1}}\dfrac{m_k(i)}{\sqrt{4\beta_k^2+m_k^2(i)}} & \textit{if } m_k(i)\omega_{ik} < m_{k-1}(i_-)\omega_{i_-,k-1}, \\ \dfrac{2}{\omega_{ik}}\dfrac{m_{k-1}(i_-)}{\sqrt{4\beta_k^2+m_{k-1}^2(i_-)}} & \textit{otherwise.} \end{cases}$$

On the appropriate interval about $\lambda_i^{(k)}$ *the second term's model, in magnitude, exceeds*

$$\begin{cases} \dfrac{2}{\omega_{ik}}\dfrac{m_{k+1}(i_+)}{\sqrt{4\beta_{k+1}^2+m_{k+1}^2(i_+)}} & \textit{if } m_{k+1}(i_+)\omega_{i_+,k+1} < m_k(i)\omega_{ik}, \\ \dfrac{2}{\omega_{i_+,k+1}}\dfrac{m_k(i)}{\sqrt{4\beta_{k+1}^2+m_k^2(i)}} & \textit{otherwise.} \end{cases}$$

Whenever one of these four expressions dominates the others, then it gives an accurate estimate of $\max|\hat{\alpha}_k'|$ *in the neighbourhood of* $\lambda_i^{(k)}$. *Otherwise, the maximum is given by a more complicated expression and it is smaller.*

Proof of (i). From §2.2 and also from (4.3),

$$\begin{aligned} \hat{\beta}_n' &= s_n\pi_n' + \pi_n s_n', \\ \hat{\beta}_n'' &= s_n\pi_n'' + s_n'\pi_n' + s_n''\pi_n. \end{aligned}$$

From Lemmas 3.1 and 3.2 and their corollaries,

$$\pi_n(\lambda_i^{(n)}) = 0 = \pi_n''(\lambda_i^{(n)}), \quad \pi_n'(\lambda_i^{(n)}) = \pm 1/\omega_{in}, \quad c_n(\lambda_i^{(n)}) = \pm\omega_{in},$$

and so,

$$|\hat{\beta}_n'(\lambda_i^{(n)})| = |s_n\pi_n'| = \left|\frac{s_n}{c_n}\right| = \sqrt{\omega_{in}^{-2} - 1}. \tag{4.13}$$

Using (4.12) gives

$$|\hat{\beta}_n''(\lambda_i^{(n)})| = 2|s_n'\pi_n'| = 2s_n \frac{|\pi_{n-1}'|}{\xi_{n-1}} \neq 0, \tag{4.14}$$

so the maximum does not occur at $\lambda_i^{(n)}$. For $\sigma \neq \lambda_i^{(n)}$ use (4.12) again to find

$$|\hat{\beta}_n'| = |s_n\pi_n' - \pi_n s_n c_n \pi_{n-1}'/\xi_{n-1}| = |s_n(\pi_n' - \pi_n\phi_{n-1})|,$$

where ϕ_{n-1} was defined in (3.18) and then modelled near $\lambda_{i_-}^{(n-1)}$ by $f_{n-1}^{(i_-)}$.

Although $\max|\phi_{n-1}| \approx 1/(2\beta_n\omega_{i,n-1})$ can be huge, we shall show that $|\hat{\beta}_n'|$ takes its local maximum near $(\lambda_i^{(n)} + \lambda_{i_-}^{(n-1)})/2$. We now analyze the model of the two factors in $\hat{\beta}_n'$.

It is convenient to abbreviate temporarily by

$$\Delta := \lambda_{i_-}^{(n-1)} - \lambda_i^{(n)}, \quad \delta := \sigma - \lambda_i^{(n)}, \quad \rho_- := \beta_n\omega_{i_-,n-1}.$$

Recall that π_{n-1} is modelled by

$$p_{n-1}^{(i_-)}(\sigma) = \frac{\pm(\sigma - \lambda_{i_-}^{(n-1)})}{\omega_{i_-,n-1}},$$

and hence $s_n = \beta_n/\sqrt{\pi_{n-1}^2 + \beta_n^2}$ is modelled by

$$\frac{\beta_n\omega_{i_-,n-1}}{\sqrt{(\delta-\Delta)^2 + \beta_n^2\omega_{i_-,n-1}^2}} = \frac{1}{\sqrt{1 + \frac{(\delta-\Delta)^2}{\rho_-^2}}}$$

on the interval of width $\omega_{i_-,n-1}m_{n-1}(i_-)$ centered at $\lambda_{i_-}^{(n-1)}$. The other factor, $\pi_n' - \pi_n\phi_{n-1}$, is modelled by $p_n^{(i)'} - p_n^{(i)}f_{n-1}^{(i_-)}$ where

$$p_n^{(i)} = \frac{(-1)^{n-i}\delta}{\omega_{\text{in}}} \quad \text{and} \quad f_{n-1}^{(i_-)} = \frac{(\delta-\Delta)}{\rho_-^2 + (\delta-\Delta)^2},$$

from (3.18) and (3.21), on the appropriate intervals centered at $\lambda_i^{(n)}$ and $\lambda_{i_-}^{(n-1)}$. Thus

$$p_n^{(i)'} - p_n^{(i)}f_{n-1}^{(i)} = \frac{(-1)^{n-i}}{\omega_{in}}\left[1 + \frac{\delta(\Delta-\delta)}{\rho_-^2 + (\Delta-\delta)^2}\right].$$

The important points are that the quantity [*] exceeds 1 only when $\delta(\Delta-\delta) > 0$, i.e., when $\sigma \in [\lambda_i^{(n)}, \lambda_i^{(n-1)}]$ and the maximum may be evaluated exactly by calculus. The details are omitted but the result is

$$\max_{\delta}|p_n^{(i)'} - p_n^{(i)}f_{n-1}^{(i_-)}| = \frac{1}{\omega_{in}}\frac{1}{2}\left(1 + \sqrt{1 + \left(\frac{\Delta}{\rho_-}\right)^2}\right)$$

and

$$\Delta - \delta = \frac{\Delta}{1 + \sqrt{1 + (\frac{\Delta}{\rho_-})^2}}.$$

The hypothesis on Δ gives

$$\frac{\Delta}{\rho_-} < \frac{100(\lambda_n^{(n)} - \lambda_1^{(n)})\omega_{in}^2}{\rho_-} = O(\omega_{in}),$$

and indicates that $|\delta| \approx |\Delta - \delta| \approx \Delta/2$, very close to the center of both intervals. Using this value in the model for s_n yields a maximum value for the model for $\hat{\beta}_n'$:

$$\frac{1}{2\omega_{in}} \frac{1 + \sqrt{1 + (\frac{\Delta}{\rho_-})^2}}{\sqrt{1 + \frac{1}{4}(\frac{\Delta}{\rho_-})^2}} = \frac{1}{\omega_{in}} \left[1 + \frac{1}{8}\left(\frac{\Delta}{\rho_-}\right)^2 + \cdots\right].$$

However, reference to (3.18) and (3.21) shows that our model for $|\pi_n' - \pi_n\phi_{n-1}|$ carries the correction factor

$$1 - \frac{3}{2}\left(\frac{\delta}{\omega_{in} m_n(i)}\right)^2 + O(\delta^3).$$

Using $|\delta| = \Delta/2$ here and combining the estimates gives

$$\max|\hat{\beta}_n'| = \frac{1}{\omega_{in}} \left[1 + \frac{1}{8}\left(\frac{\Delta}{\rho_-}\right)^2 - \frac{3}{8}\left(\frac{\Delta}{\omega_{in} m_n(i)}\right)^2 + O(\Delta^3)\right]. \qquad \square$$

Proof of (ii). For all σ and $2 \leq k < n$,

$$\hat{\beta}_k' = \hat{\beta}_k(\phi_k - \phi_{k-1}) = \beta_k \frac{\xi_k}{\xi_{k-1}}(\phi_k - \phi_{k-1}),$$

where

$$\xi_k = \sqrt{\pi_k^2 + \beta_{k+1}^2} \quad \text{(see (2.9))},$$

$$\phi_k := \frac{\pi_k \pi_k'}{\xi_k^2} :\approx f_k^{(i)} = \frac{\sigma - \lambda_i^{(k)}}{\beta_{k+1}^2 \omega_{ik}^2 + (\sigma - \lambda_i^{(k)})^2} \quad \text{(see (3.21))}.$$

The interval associated with $f_k^{(i)}$ is

$$[\lambda_i^{(k)} - m_k(i)\omega_{ik}/2, \lambda_i^{(k)} + m_k(i)\omega_{ik}/2].$$

To analyze the model for $\hat{\beta}_k'$ it is convenient to abbreviate as follows:

$$\delta := \sigma - \lambda_i^{(k)}, \quad \Delta := \lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}, \quad \rho_k := \beta_{k+1}\omega_{ik}, \quad \rho_{k-1} := \beta_k \omega_{i_-,k-1}.$$

$\hat{\beta}_k'$ is modelled by a difference of two functions:

$$\hat{\beta}_k \left[f_k^{(i)} - f_{k-1}^{(i_-)}\right].$$

We consider $\beta_k \xi_k f_k^{(i)}/\xi_{k-1}$ first. By calculus we find that if $\beta_{k+1} < m_k(i)/2$, then $|f_k^{(i)}|$ attains its global maximum value $1/2\rho_k$ at $\delta = \pm\rho_k$ within the associated interval. At these points,

$$\begin{aligned}\frac{\xi_k}{\xi_{k-1}} &\approx \left[\frac{\rho_k^2/\omega_{ik}^2 + \beta_{k+1}^2}{(\rho_k - \Delta)^2/\omega_{i_-,k-1}^2 + \beta_k^2}\right]^{1/2} \\ &= \frac{\beta_{k+1}}{\beta_k}\left[\frac{2}{1 + (\rho_k/\rho_{k-1})^2}\right]^{1/2} [1 + O(\omega_{ik})] .\end{aligned}$$

Next consider $\beta_k \xi_k f_{k-1}^{(i_-)}/\xi_{k-1}$. By calculus we find that if $\beta_k < m_{k-1}(i_-)/2$, then $|f_{k-1}^{(i_-)}|$ attains its global maximum value $1/2\rho_{k-1}$, within its associated interval, at $\delta - \Delta = \pm\rho_{k-1}$. At these points,

$$\begin{aligned}\frac{\xi_k}{\xi_{k-1}} &\approx \left[\frac{(\rho_{k-1} \pm \Delta)^2/\omega_{ik}^2 + \beta_{k+1}^2}{\rho_{k-1}^2/\omega_{i_-,k-1}^2 + \beta_k^2}\right]^{1/2} \\ &= \frac{\beta_{k\pm1}}{\beta_k}\left[\frac{1 + (\rho_{k-1}/\rho_k)^2}{2}\right]^{1/2} [1 + O(\omega_{ik})] .\end{aligned}$$

If, however, $\beta_{k+1} > m_k(i)/2$, then the maximum of $|f_k^{(i)}|$ on its associated interval is approximately $m_k(i)/(2\beta_{k+1})$ of its global maximum $1/(2\rho_k)$ and similarly for $|f_{k-1}^{(i_-)}|$ when $\beta_k > m_{k-1}(i_-)/2$. We will not consider these two cases further because they yield more complicated expressions and smaller maxima.

Returning to the situation when $|f_k^{(i)}|$ and $|f_{k-1}^{(i_-)}|$ attain their global maxima within their appropriate intervals, as shown in Fig. 2, we distinguish two cases.

(a) $\rho_k = \beta_{k+1}\omega_{ik} < \rho_{k-1} = \beta_k \omega_{i_-,k-1}$. The first expression shows that

$$\left|\frac{\xi_k}{\xi_{k-1}}\right|_{\delta=\rho_k} > \frac{\beta_{k+1}}{\beta_k}(1 + O(\omega_{ik})).$$

Thus the first term's model $\hat{\beta}_k f_k^{(i)}$ satisfies

$$|\hat{\beta}_k f_k^{(i)}(\lambda_i^{(k)} \pm \rho_k)| > \beta_k \frac{\beta_{k+1}}{\beta_k}\frac{1}{2\rho_k}(1 + O(\omega_{ik})) = \frac{1}{2\omega_{ik}}(1 + O(\omega_{ik})).$$

(b) $\rho_{k-1} < \rho_k$. The second expression shows that

$$\left|\frac{\xi_k}{\xi_{k-1}}\right|_{|\delta-\Delta|=\rho_{k-1}} > \frac{\beta_{k+1}}{\sqrt{2}\beta_k}(1 + O(\omega_{ik})).$$

Thus the second term's model $\hat{\beta}_k f_{k-1}^{(i_-)}$ satisfies

$$\begin{aligned}|\hat{\beta}_k f_{k-1}^{(i_-)}(\lambda_i^{(k)} - \Delta \pm \rho_{k-1})| &> \beta_k \frac{\beta_{k+1}}{\sqrt{2}\beta_k}\frac{1}{2\rho_{k-1}}(1 + O(\omega_{ik})) \\ &= \frac{\beta_{k+1}}{\sqrt{2}\beta_k}\frac{1}{2\omega_{i_-,k-1}}(1 + O(\omega_{ik})).\end{aligned}$$

When one of the two terms derived in cases (a) and (b) dominates the other (by at least a factor of 2) then the larger term gives an accurate enough estimate of the maximum magnitude of the model and of $\hat{\beta}_k'$. By the assumption $\Delta := \lambda_{i_-}^{(k-1)} - \lambda_i^{(k)} = O(\omega_{ik}^2)$, $f_k^{(i)}$ and $f_{k-1}^{(i_-)}$ have the same sign at their global maxima and so cause cancellation in the expression for $\hat{\beta}_k'$. When any of our conditions fail, either $\beta_{k+1} > m_k(i)/2$ or $\beta_k > m_{k-1}(i_-)/2$, or when $\rho_{k-1} \approx \rho_k$, then the f functions have smaller values and/or there is a significant cancellation between them. In such cases $|\hat{\beta}_k'|$ need not be

$$O\left(\frac{1}{\min\left\{\omega_{ik}, \omega_{i_-,k-1}\right\}}\right) \quad \text{near } \lambda_i^{(k)}. \qquad \square$$

Remark. Figures 2 and 3 illustrate the preceding analysis. The matrix is W_{17}^-. $k = 8$, $i = 5$. Then $k-1 = 7$ and $i_- = 4$ by its definition in (*) preceding Theorem 4.2. Note that $\omega_{ik} = 0.032$, $\omega_{i_-,k-1} = 0.116$, $\rho_i = \beta_{k+1}\omega_{ik}/\beta_k\omega_{i_-,k-1} = 0.27$, $m_k(i) = 3.337$, $m_{k-1}(i_-) = 2.31$, $|\Delta| = 0.0037$. Since $\beta_{k+1} = \beta_k = 1$, the chosen evaluation points are within the domain of the model. The first graph shows ϕ_k and ϕ_{k-1} and the second shows $\hat{\beta}_k'$ and the model

$$\text{(4.15)} \qquad \beta_k \left(\frac{\beta_{k+1}^2 + \delta^2/\omega_{ik}^2}{\beta_k^2 + (\delta - \Delta)^2/\omega_{i_-,k-1}^2}\right)^{1/2} (f_k^{(i)} - f_{k-1}^{(i_-)}).$$

Proof of (iii). From the proof of Theorem 4.1 we know that $\hat{\alpha}_n' = 2(\pi_n' c_n + 1) = 2\pi_n c_n'$. Thus $\hat{\alpha}_n'(\lambda_{i_-}^{(n-1)})$=2 since c_n vanishes on the spectrum of T_{n-1}, and $\hat{\alpha}_n'(\lambda_i^{(n)})$=0 since π_n vanishes on the spectrum of T_n. However, $|\hat{\alpha}_n'|$ can attain large values when σ is close to these points but *not* between them. The model $p_n^{(i)}$ for π_n is valid when $|\sigma - \lambda_i^{(n)}| < m_n(i)\omega_{in}/2$ and the model

$$p_{n-1}^{(i-1)}/\sqrt{\beta_n^2 + (p_{n-1}^{(i-1)})^2}$$

for c_n is valid for $|\delta| := |\sigma - \lambda_{i_-}^{(n-1)}| < m_{n-1}(i_-)\omega_{i_-,n-1}/2$. On the intersection of these intervals the model for $\hat{\alpha}_n'$ is

$$\begin{aligned} 2\left(p_n^{(i)\prime} p_{n-1}^{(i-1)} \Big/ \sqrt{\beta_n^2 + (p_{n-1}^{(i-1)})^2} + 1\right) &= 2\left[\frac{(-1)^i}{\omega_{in}} \frac{(-1)^{i_-}\delta}{\omega_{i_-,n-1}} \Big/ \frac{(\beta_n^2\omega_{i_-,n-1}^2 + \delta^2)^{1/2}}{\omega_{i_-,n-1}} + 1\right] \\ &= 2\left[\frac{(-1)^{i+i_-}\delta}{\omega_{in}(\beta_n^2\omega_{i_-,n-1}^2 + \delta^2)^{1/2}} + 1\right]. \end{aligned}$$

The function $\delta/(\tau^2 + \delta^2)^{1/2}$ is a monotone-increasing function for all δ, and so the model takes its maximum magnitude at the boundary of the associated interval, namely,

$$\begin{aligned} &2\left[\frac{m_{n-1}(i_-)}{\omega_{in}(4\beta_n^2 + m_{n-1}^2(i_-))^{1/2}} + 1\right] \quad \text{if } m_{n-1}(i_-)\omega_{i_-,n-1} < m_n(i)\omega_{in}, \\ &2\left[\frac{m_n(i)}{\omega_{i_-,n-1}(4\beta_n^2 + m_n^2(i))^{1/2}} + 1\right] \quad \text{otherwise.} \end{aligned}$$

We evaluate at the left end if $i + i_-$ is odd. The evaluation point

$$\sigma = \pm\frac{1}{2}m_n(i)\omega_{in} - \lambda_{i_-}^{(n-1)} = \pm\frac{1}{2}m_n(i)\omega_{in} - \lambda_i^{(n)} + [\lambda_i^{(n)} - \lambda_{i_-}^{(n-1)}]$$

is just outside the standard interval in some cases but the excess is $O(\omega_{in}^2)$ and permits a cleaner expression for the model. □

Proof of (iv). From (4.0) and (4.2),

$$\begin{aligned}
\hat{\alpha}'_k &= \gamma'_k - \gamma'_{k+1} \\
&= \pi_k c'_k + \pi'_k c_k - (\pi_{k+1} c'_{k+1} + \pi'_{k+1} c_{k+1}) \\
&= \begin{cases} 2\pi_k c'_k - 2\pi_{k+1} c'_{k+1} \\ 2\pi'_k c_k - 2\pi'_{k+1} c_{k+1} \end{cases} \quad \text{by (4.5)} \\
&= 2\left(\frac{\pi'_k \pi_{k-1}}{\xi_{k-1}} - \frac{\pi'_{k+1}\pi_k}{\xi_k} \right) \quad \text{by (2.10).}
\end{aligned}$$

Near $\lambda_i^{(k)}$, $\hat{\alpha}'_k$ is modelled by

$$\frac{2p_k^{(i)'} p_{k-1}^{(i_-)}}{(\beta_k^2 + (p_{k-1}^{(i_-)})^2)^{1/2}} - \frac{2p_{k+1}^{(i^+)'} p_k^{(i)}}{(\beta_{k+1}^2 + (p_k^{(i)})^2)^{1/2}}.$$

Here is the difference of two terms of the type analyzed in the proof of (iii). Each term is monotonic on the appropriate interval. The first term is valid on the intersection of the intervals

$$\begin{aligned}
&\text{center } \lambda_i^{(k)}, && \text{radius } \tau_k := \tfrac{1}{2} m_k(i)\omega_{ik}, \\
&\text{center } \lambda_{i_-}^{(k-1)}, && \text{radius } \tau_{k-1} := \tfrac{1}{2} m_{k-1}(i_-)\omega_{i_-,k-1},
\end{aligned}$$

and the second term is valid on the intervals

$$\begin{aligned}
&\text{center } \lambda_i^{(k)}, && \text{radius } \tau_k := \tfrac{1}{2} m_k(i)\omega_{ik}, \\
&\text{center } \lambda_{i_+}^{(k+1)}, && \text{radius } \tau_{k+1} := \tfrac{1}{2} m_{k+1}(i_+)\omega_{i_+,k+1}.
\end{aligned}$$

From the analysis of $\hat{\alpha}'_n$, the magnitude of the first term achieves

$$\begin{cases} \dfrac{2}{\omega_{i_-,k-1}} \dfrac{m_k(i)}{(4\beta_k^2 + m_k^2(i))^{1/2}} & \text{if } \tau_k < \tau_{k-1}, \\[2ex] \dfrac{2}{\omega_{ik}} \dfrac{m_{k-1}(i_-)}{(4\beta_k^2 + m_{k-1}^2(i_-))^{1/2}} & \text{if } \tau_{k-1} < \tau_k. \end{cases}$$

Similarly, the second term achieves

$$\begin{cases} \dfrac{2}{\omega_{ik}} \dfrac{m_{k+1}(i_+)}{(4\beta_{k+1}^2 + m_{k+1}^2(i_+))^{1/2}} & \text{if } \tau_{k+1} < \tau_k, \\[2ex] \dfrac{2}{\omega_{i_+,k+1}} \dfrac{m_k(i)}{(4\beta_{k+1}^2 + m_k^2(i))^{1/2}} & \text{if } \tau_k < \tau_{k+1}. \end{cases}$$

□

Before we leave this section, we give Theorem 4.4, whose proof appears in [3].

THEOREM 4.4. *Suppose that T is unreduced. Then*

$$\omega_{ik}^2 < \begin{cases} \dfrac{\lambda_1^{(k-1)} - \lambda_1^{(k)}}{\lambda_2^{(k)} - \lambda_1^{(k)}}, & i = 1, \\[2ex] \dfrac{\lambda_i^{(k)} - \lambda_{i-1}^{(k-1)}}{\lambda_i^{(k)} - \lambda_{i-1}^{(k)}} \cdot \dfrac{\lambda_i^{(k-1)} - \lambda_i^{(k)}}{\lambda_{i+1}^{(k)} - \lambda_i^{(k)}}, & i \neq 1, k, \\[2ex] \dfrac{\lambda_k^{(k)} - \lambda_{k-1}^{(k-1)}}{\lambda_k^{(k)} - \lambda_{k-1}^{(k)}}, & i = k, \end{cases}$$

and

$$\omega_{ik}^2 > \begin{cases} \dfrac{\lambda_1^{(k-1)} - \lambda_1^{(k)}}{\lambda_k^{(k)} - \lambda_1^{(k)}}, & i = 1, \\[2ex] \dfrac{\lambda_i^{(k)} - \lambda_{i-1}^{(k-1)}}{\lambda_i^{(k)} - \lambda_1^{(k)}} \cdot \dfrac{\lambda_i^{(k-1)} - \lambda_i^{(k)}}{\lambda_k^{(k)} - \lambda_i^{(k)}}, & i \neq 1, k, \\[2ex] \dfrac{\lambda_k^{(k)} - \lambda_{k-1}^{(k-1)}}{\lambda_k^{(k)} - \lambda_1^{(k)}}, & i = k. \end{cases}$$

Proof. See [3]. □

5. Premature deflation and the monitoring algorithm.

The reader is referred to (2.17) for a picture of the active submatrix of the intermediate matrix $T^{(k)}$ that occurs in the transformation of $T = T^{(1)}$ to $\hat{T} = T^{(n)}$.

Observe that if $s_k\pi_k$ and $\beta_{k+1}c_k$ are replaced by zero to give $\dot{T}^{(k)}$, then $\dot{T}^{(k)}$ must exhibit an eigenvalue in the (k,k) position, and if row and column k are deleted, then the new $(n-1) \times (n-1)$ matrix is tridiagonal.

When $s_k\pi_k$ and $\beta_{k+1}c_k$ are small enough and the (k,k) entry equals the shift σ then we say that *premature deflation* occurred at step k in the implicit shift version of the QR transform.

The next result implies that a negligible value of some ω_{ik}^2 (compared to 1) is a necessary condition for premature deflation.

LEMMA 5.1. *On the interval* $\left[\lambda_i^{(k)}, \lambda_{i_-}^{(k-1)}\right]$

$$s_k^2\pi_k^2 + \beta_{k+1}^2c_k^2 \geq 2\left(\frac{\beta_{k+1}^2\omega_{ik}^2}{\beta_k^2\omega_{i_-,k-1}^2 + \beta_{k+1}^2\omega_{i,k}^2}\right)\left(\Psi_i^{(k)}\omega_{ik}\right)^2\left(1 + O\left[\left(\frac{\Delta_i}{\beta_k\omega_{i_-,k-1}}\right)^2\right]\right),$$

where

$$\Psi_i^{(k)} = \begin{cases} \lambda_2^{(k)} - \lambda_1^{(k)}, & i = 1, \\[1ex] \dfrac{\left(\lambda_{i+1}^{(k)} - \lambda_i^{(k)}\right)\left(\lambda_i^{(k)} - \lambda_{i-1}^{(k)}\right)}{|\lambda_i^{(k)} - \lambda_i^{(k-1)}|}, & i \neq 1, k, \\[1ex] \lambda_k^{(k)} - \lambda_{k-1}^{(k)}, & i = k; \end{cases}$$

$$\Delta_i^{(k)} = |\lambda_i^{(k)} - \lambda_{i_-}^{(k-1)}|.$$

Proof. Since $\Delta_i^{(k)} = O(\omega_{ik}^2)$ the linear models for π_k and π_{k-1} are valid. Write $\delta = \sigma - \lambda_i^{(k)}, \omega = \omega_{ik}, \tilde{\omega} = \omega_{i_-,k-1}$, and observe that

$$
\begin{aligned}
s_k^2\pi_k^2 + \beta_{k+1}^2 c_k^2 &= \frac{\frac{\beta_k^2\delta^2}{\omega^2} + \frac{\beta_{k+1}^2(\widetilde{\delta-\Delta})^2}{\omega}^{2}}{\beta_k^2 + (\delta-\Delta)^2/\tilde{\omega}^2} \\
&= \frac{\tilde{\omega}^2\beta_k^2\delta^2 + \omega^2\beta_{k+1}^2(\Delta-\delta)^2}{\beta_k^2\omega^2\tilde{\omega}^2}\left(1 + O\left(\frac{\Delta}{\beta_k\tilde{\omega}}\right)^2\right) \\
&> \frac{2\tilde{\omega}^2\beta_k^2\omega^2\beta_{k+1}^2\Delta^2}{(\tilde{\omega}^2\beta_k^2 + \omega^2\beta_{k+1}^2)\beta_k^2\omega^2\tilde{\omega}^2}\left(1 + O\left(\frac{\Delta}{\beta_k\tilde{\omega}}\right)^2\right) \\
&\quad \text{(the minimum for } \delta \in [0,\Delta]) \\
&= 2\frac{\beta_{k+1}^2}{(\tilde{\omega}^2\beta_k^2 + \omega^2\beta_{k+1}^2)}\Delta^2\left(1 + O\left(\frac{\Delta}{\beta_k\tilde{\omega}}\right)^2\right).
\end{aligned}
$$

By Theorem 4.4, $\Delta_i^{(k)} > \Psi_i^{(k)}\omega_{ik}^2$, the result follows. Note that the term ω_{ik}^4 has been rearranged in the lemma's statement. □

Note that at $\sigma = \lambda_i^{(k)}$,

$$
s_k^2\pi_k^2 + \beta_{k+1}^2 c_k^2 = \beta_{k+1}^2\omega_{ik}^2
$$

since $\pi_k = 0$, and outside $[\lambda_i^{(k)}, \lambda_{i_-}^{(k-1)}]$ the sum rises rapidly to $O(\text{spread}_k^2)$.

COROLLARY. *Premature deflation occurs if and only if σ is very close to $\lambda_i^{(k)}$ and ω_{ik}^2 is negligible. By neglecting the entries $s_k\pi_k$ and $\beta_{k+1}c_k$ and deleting the kth row and column the changes made in the eigenvalues are bounded by $(s_k^2\pi_k^2 + \beta_{k+1}^2c_k^2)/\min(|\sigma-\lambda_{i-1}^{(k)}|, |\sigma-\lambda_{i+1}^{(k)}|)$ and the error angles in the eigenvectors are bounded by $(s_k^2\pi_k^2 + \beta_{k+1}^2c_k^2)^{1/2}/\min(|\sigma-\lambda_{i-1}^{(k)}|, |\sigma-\lambda_{i+1}^{(k)}|)$.*

Thus premature deflation at step k accompanies great sensitivity of $\hat{\beta}_k, \hat{\beta}_{k+1}$, and $\hat{\alpha}_{k+1}$ to small changes in σ. In many, but not all, cases a tiny value of ω_{ik} will be associated with a tiny value of $\omega_{i_+,k+1}$.

The pattern is as follows. If ω_{ik} is tiny then $y_k^{(n)}$, T_k's eigenvector for $\lambda_i^{(k)}$ with zero entries appended, will be an excellent approximate eigenvector for all extensions to T_k. The only way that subsequent values ω_{in} can increase is by the presence of two or more eigenvalues of T_n close to $\lambda_i^{(k)}$. In such a case, $y_k^{(n)}$ is a linear combination of the eigenvectors belonging to all $\lambda_j^{(n)}$ close to $\lambda_i^{(k)}$ but not close to any of others. In fact, $y_k^{(n)}$ is usually close to a bisector of a pair of eigenvectors, and their last entries will be nearly equal and not necessarily small.

Here is an illustration of this phenomenon.

Example 5.1.

$$
T = \text{tridiag}\begin{pmatrix} & 1 & 1\cdots 1 & 1 & \\ 13 & 0 & \cdots & 0 & 13 \\ & 1 & 1\cdots 1 & 1 & \end{pmatrix},
$$

$$
\alpha_i = 0, (i = 2,\ldots,24), \quad \alpha_1 = \alpha_{25} = 13, \quad \beta_i = 1, (i = 2,\ldots,25).
$$

This matrix has two close eigenvalues. The instability occurs midway through the QR transform with shift at one of these two eigenvalues. Rather than exhibit matrices of

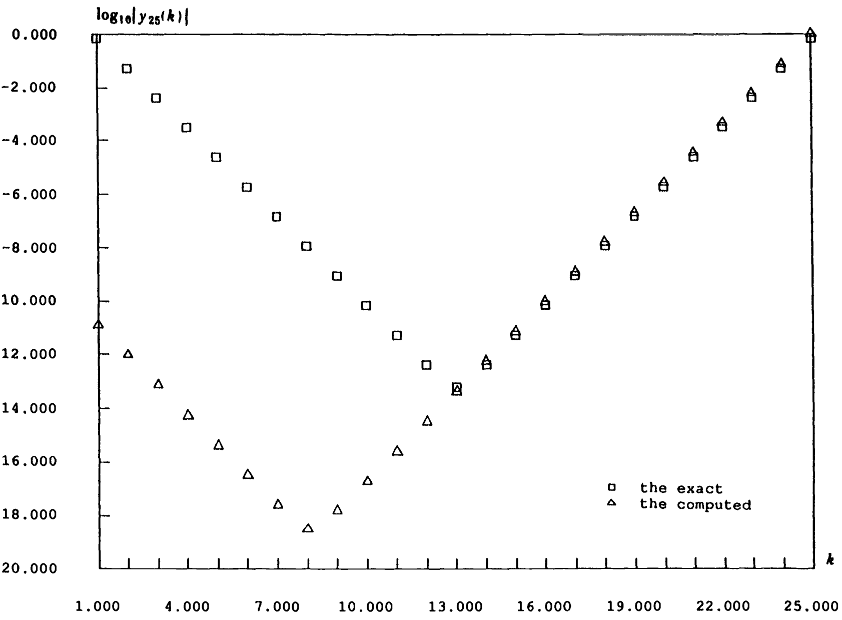


FIG. 8. *Vector components of y_{25} defined in* (2.2), *in Example* 5.1.

this size, we plot in Fig. 8 the exact and the computed vectors y_{25} on a logarithmic scale. At first sight it is surprising that the top half of y is quite wrong, while the lower half seems reasonable. However, reference to (2.2) reveals that late bad values of s all propagate to the top of y. Further inspection of the exact and computed c and s values shows that the first eight are good. Then instability sets in, but the last two rows are only wrong by a factor of 2. So the lower half of y only looks reasonable on a logarithmic scale; it has barely one bit of accuracy.

5.1. Forward stability for suitable shift strategies. Our analysis has focused on the role of the numbers ω_{ik} in one QR transform. In the QR algorithm a sequence of QR transformations is applied to the original matrix with carefully chosen shifts. At each transformation the eigenvectors change and the ω_{ik}'s change along with them. In the last two steps of the algorithm the last off-diagonal entry diminishes rapidly. To see the effect on the ω_{in} it suffices to consider an extreme case. Suppose that $\beta_n = 0$ but the algorithm does not notice this fact. Any reasonable shift strategy will force $\sigma = \alpha_n$ in this case, and the associated eigenvector is e_n, with its $\omega_{in} = 1$. All the other values of ω_{in} vanish since $\sum_i \omega_{in}^2 = 1$. The analysis of the previous section shows that the QR transform with shift α_n is stable, but for any shift at an eigenvalue of T_{n-1} (not T_n), the last plane rotation is arbitrary and $\hat{\beta}_n$ is undefined. In this case, the QR algorithm with any sensible shift strategy is forward stable, although the QR transform with "wrong" shifts is completely unstable.

However, forward instability can still occur in the standard QR algorithm, but only when the shift is very close to a cluster of eigenvalues equal to working precision. This was shown in Example 2.4 of §2.3.

5.2. The ultimate shift strategy. The QR transformation of an $n \times n$ real symmetric tridiagonal matrix requires about $10n$ flops. The most expensive feature of a QR algorithm that produces eigenvectors is the accumulation of all the plane rotations. This is an $O(n^2)$ process for each transform, and the cost is directly proportional to the number of QR transformations.

These facts suggested the use of a two-phase process. First, compute the eigenvalues by any means, keeping a copy of the original tridiagonal. Second, apply the QR transformation using the eigenvalues as shifts (see [4, p. 164]).

Our analysis shows that this strategy invites forward instability. In fact, it will occur for each eigenvalue whose normalized eigenvector has a tiny bottom element.

This is the first reason we have seen for being cautious about the use of the ultimate shift strategy. However, the simple modification of the QR transform given below can preserve forward stability.

5.3. QR with monitoring. The idea of the algorithm is to check for premature deflation and stop the QR transformation as soon as it occurs. Then remove the row and column with the isolated eigenvalue. This is not a pure QR algorithm, but it does use plane rotations and does restore tridiagonal form.

Since the monitoring test usually fails, it is preferable to break it into two parts so that the part that is always made involves no arithmetic operations.

```
If $|\pi_k| < \sqrt{\epsilon}\|T\|/\sqrt{n}$ then
    If $|\pi_k| + \beta_{k+1}|c_k| < \sqrt{\epsilon}\ (|\alpha_{k-1}| + |\alpha_k| + |\alpha_{k+1}| + \beta_{k-1} + \beta_k)$ then
        pull up the remaining entries of $T$ to overwrite row and column $k$.
```

Such an algorithm has been used regularly for deflation purposes in the context of the Lanczos algorithm. Since a copy of the undeflated T matrix is preserved for the eigenvector computations at the end of the Lanczos run there is no harm in suppressing off-diagonal entries as large as $\sqrt{\epsilon}\|T\|$ in the deflated T. By the Corollary to Lemma 5.1, such modifications could indeed induce a $\sqrt{\epsilon}$ twist in T's eigenvectors. Whether such alterations damage the Ritz vectors (the approximate eigenvectors of the operator driving the Lanczos algorithm) is not clear.

6. Conclusion. The occasional forward instability of the QR transformation is not a well-known phenomenon and so we have tried to give a full account of it. In particular, we have shown its intimate connection with premature deflation of the shift and with the quantities ω_{ik}. Since the $\{\omega_{ik}\}$ change at each step in the QR algorithm, because the eigenvectors change, forward instability is unlikely to occur, although, by Example 2.4, it can happen.

On the other hand, forward instability is seen quite frequently when deflating stabilized eigenvalues in the Lanczos algorithm and sometimes when using the ultimate shift strategy with the QR algorithm.

Our analysis of the function $\pi_k(\sigma)$ and of the derivative of the tridiagonal QR transform are full of details that will burden the reader. To put this aspect in perspective, we would like to say that the entries in $\hat{T}$ are complicated functions and it is far from obvious where their derivatives peak. We could see no other way than direct elucidation of $\max_\sigma |\hat{\beta}_k'|$ and $\max_\sigma |\hat{\alpha}_k'|$ to establish that small values of some ω_{ik} are generically sufficient as well as necessary for the occurrence of forward instability.

Acknowledgement. The authors are grateful for the extremely careful and helpful reviewing by the two referees and the editor Linda Kaufman throughout several revisions of this work.

REFERENCES

[1] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[2] W. B. Gragg and W. J. Harrod, *The numerically stable reconstruction of Jacobi matrices from spectral data*, Numer. Math., 44 (1984), pp. 317–336.

[3] R. O. Hill and B. N. Parlett, *Refined interlacing properties*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 239–247.

[4] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[5] B. T. Smith, J. M. Boyle, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, *Matrix Eigensystem Routines*—EISPACK *Guide*, in Lecture Notes in Computer Science, 2nd Ed., Vol. 6, Springer-Verlag, New York, 1976.

[6] G. W. Stewart, *Incorporating origin shifts into the* QR *algorithm for symmetric tridiagonal matrices*, Comm. Assoc. Comput. Mach., 13 (1970), pp. 365–367.

[7] ———, *Perturbation bounds for the* QR *factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.

[8] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

[9] ———, *The calculation of the eigenvectors of codiagonal matrices*, Comput. J., 1 (1958), pp. 90–96.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 317–333, April 1993

001

OPTIMIZATION BY DIRECT SEARCH IN MATRIX COMPUTATIONS*

NICHOLAS J. HIGHAM†

Abstract. A direct search method attempts to maximize a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ using function values only. Many questions about the stability and accuracy of algorithms in matrix computations can be expressed in terms of the maximum value of some easily computable function f. For a variety of algorithms it is shown that direct search is capable of revealing instability or poor performance, even when such failure is difficult to discover using theoretical analysis or numerical tests with random or nonrandom data. Informative numerical examples generated by direct search provide the impetus for further analysis and improvement of an algorithm. The direct search methods used are the method of alternating directions and the multi-directional search method of Dennis and Torczon. The problems examined include the reliability of matrix condition number estimators and the stability of Strassen's fast matrix inversion method.

Key words. optimization, matrix computations, direct search method, numerical stability, Gaussian elimination, matrix condition number estimation, fast matrix multiplication, Vandermonde system, matrix inverse

AMS(MOS) subject classifications. 65F05, 65G05, 65K10

1. Introduction. Is Algorithm X numerically stable? How large can the growth factor be for Gaussian elimination with pivoting strategy P? By how much can condition estimator C underestimate the condition number of a matrix? These types of questions are fundamental in the analysis of algorithms in matrix computations. Usually, one attempts to answer such questions by a combination of theoretical analysis and numerical experiments with random and nonrandom data. In this work we show that a third approach can be a valuable supplement to the first two: phrase the question as an optimization problem and apply a direct search method.

A direct search method for the problem

$$\max_{x \in \mathbb{R}^n} f(x), \qquad f : \mathbb{R}^n \rightarrow \mathbb{R} \tag{1.1}$$

is a numerical method that attempts to locate a maximizing point using function values only, and which does not attempt to estimate derivatives of f. Such methods are usually based on heuristics that do not involve assumptions about the function f. Various direct search methods have been developed; for surveys, see [43], [52], and [53]. Most of these methods were developed in the 1960s, in the early days of numerical optimization. For problems in which f is smooth, direct search methods have largely been supplanted by more sophisticated optimization methods that use derivatives (such as quasi-Newton methods and conjugate gradient methods), but they continue to find use in applications where f is not differentiable, or even not continuous. These applications range from chemical analysis [46], where direct search methods have found considerable use, to the determination of drug doses in the treatment of cancer [4]; in both applications the evaluation of f is affected by experimental errors. Lack of smoothness of f, and the difficulty of obtaining derivatives when they exist, are characteristic of the optimization problems we consider here.

* Received by the editors February 25, 1991; accepted for publication (in revised form) October 15, 1991.

† Department of Mathematics, University of Manchester, Manchester, M13 9PL, United Kingdom (na.nhigham@na-net.ornl.gov).

Our aims and techniques can be illustrated using the example of Gaussian elimination (GE). Wilkinson's classic backward error analysis [60] shows that the stability of the process for $A \in \mathbb{R}^{n\times n}$ is determined by the size of the growth factor

$$\rho_n(A) = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|},$$

where the $a_{ij}^{(k)}$ are the intermediate elements generated during the elimination. For a given pivoting strategy we would therefore like to know how big $\rho_n(A)$ can be. To obtain an optimization problem of the form (1.1) we let $x = \text{vec}(A) \in \mathbb{R}^{n^2}$, where vec$(A)$ comprises the columns of A strung out into one long vector, and we define $f(x) = \rho_n(A)$. Then we wish to determine

$$\max_{x\in\mathbb{R}^{n^2}} f(x) \equiv \max_{A\in\mathbb{R}^{n\times n}} \rho_n(A).$$

Suppose, first, that no pivoting is done. Then f is defined and continuous at all points where the elimination does not break down, and it is differentiable except at points where there is a tie for the maximum in the numerator or denominator of the expression defining $\rho_n(A)$. We took $n = 4$ and applied the direct search maximizer MDS (described in §3) to $f(x)$, starting with the identity matrix $A = I_4$. After 11 iterations and 433 function evaluations, the maximizer converged,[1] having located the matrix[2]

$$B = \begin{bmatrix} 1.7846 & -0.2760 & -0.2760 & -0.2760 \\ -3.3848 & 0.7240 & -0.3492 & -0.2760 \\ -0.2760 & -0.2760 & 1.4311 & -0.2760 \\ -0.2760 & -0.2760 & -0.2760 & 0.7240 \end{bmatrix},$$

for which $\rho_4(B) = 1.23 \times 10^5$. (The large growth is a consequence of the submatrix $B(1\colon 3, 1\colon 3)$ being ill conditioned; B itself is well conditioned.) Thus the optimizer readily shows that $\rho_n(A)$ can be very large for GE without pivoting.

Next, consider GE with partial pivoting. Here, at the kth stage of the elimination, rows are interchanged so that $|a_{kk}^{(k)}| \geq |a_{ik}^{(k)}|$, $i = k\colon n$. Now f is defined everywhere but is usually discontinuous when there is a tie in the choice of pivot element, because then an arbitrarily small change in A can alter the pivot sequence. We applied the maximizer MDS to f, this time starting with the orthogonal matrix $A \in \mathbb{R}^{4\times 4}$ with $a_{ij} = (2/\sqrt{2n+1})\sin(2ij\pi/(2n+1))$ [34], for which $\rho_4(A) = 2.32$. After 29 iterations and 1169 function evaluations the maximizer converged to a matrix B with $\rho_4(B) = 5.86$. We used this matrix to start the maximizer AD (described in §3); it took five iterations and 403 function evaluations to converge to the matrix

$$C = \begin{bmatrix} 0.7248 & 0.7510 & 0.5241 & 0.7510 \\ 0.7317 & 0.1889 & 0.0227 & -0.7510 \\ 0.7298 & -0.3756 & 0.1150 & 0.7511 \\ -0.6993 & -0.7444 & 0.6647 & -0.7500 \end{bmatrix},$$

[1] In the optimizations of this section we used the convergence tests described in §3 with tol $= 10^{-3}$.

[2] All numbers quoted are rounded to the number of significant figures shown.

for which $\rho_4(C) = 7.939$. Note that this matrix is not of the form

$$A = \begin{bmatrix} 1 & & & 1 \\ -1 & 1 & & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{bmatrix}$$

identified by Wilkinson [60] as yielding the maximum possible growth $\rho_n = 2^{n-1}$ for partial pivoting. The whole set of matrices $A \in \mathbb{R}^{n\times n}$ for which $\rho_n(A) = 2^{n-1}$ is described in [34], and C is one of these matrices, modulo the convergence tolerance.

These examples, and others presented below, illustrate the following attractions of using direct search methods to aid the understanding of algorithms in matrix computations.

(1) The simplest possible formulation of optimization problem is often sufficient to yield useful results. Derivatives are not needed, and direct search methods tend to be insensitive to lack of smoothness in the objective function f. Unboundedness of f is a favourable property—direct search methods usually quickly locate large values of f.

(2) Good progress can often be made from simple starting values, such as an identity matrix. However, prior knowledge of the problem may provide a good starting value that can be substantially improved (as in the partial pivoting example).

(3) Usually it is the global maximum of f in (1.1) that is desired (although it is often sufficient to know that f can exceed a specified value). When a direct search method converges it will, in general, at best have located a *local* maximum—and in practice the maximizer may simply have stagnated, particularly if a slack convergence tolerance is used. However, further progress can often be made by *restarting* the same (or a different) maximizer, as in the partial pivoting example. This is because for methods that employ a simplex (such as the MDS method), the behaviour of the method starting at x_0 is determined not just by x_0 but also by the $n+1$ vectors in the initial simplex constructed at x_0.

(4) The numerical information revealed by direct search provides a starting point for further theoretical analysis. For example, the GE experiments above strongly suggest the (well-known) results that $\rho_n(A)$ is unbounded without pivoting and bounded by 2^{n-1} for partial pivoting, and inspection of the numerical data suggests the methods of proof.

When applied to smooth problems the main disadvantages of direct search methods are that they have at best a linear rate of convergence and they are unable to determine the nature of the point at which they terminate (since derivatives are not calculated). These disadvantages are less significant for the problems we consider, where it is not necessary to locate a maximum to high accuracy and objective functions are usually nonsmooth. (Note that these disadvantages are not necessarily shared by methods that implicitly or explicitly *estimate* derivatives using function values, such as methods based on conjugate directions [43], [44]; however, these are not normally regarded as direct search methods.)

The rest of this paper is organized as follows. In §2 we summarize related work and explain what is new about our approach. In §3 we describe the alternating directions (AD) method and the multidirectional search (MDS) method that we have used in this work. All our experiments were done using the interactive package MATLAB [40]. We used 80286 and 80386 PC-compatible machines that are of similar overall speed to a Sun 3/50 workstation. Most of the optimization runs that we describe took less than an hour of computing time.

In §4 we show how direct search methods can provide insight into the performance of matrix condition number estimators, for which the construction of "counterexamples" is usually difficult. In §5 we describe some other, miscellaneous problems in matrix computations that can be successfully explored using direct search. Finally, in §6, we offer some conclusions.

2. Related work. This work was inspired by two papers published in the 1970s by Miller [37], [38] and by recent work of Rowan [47]. Miller's papers describe a way of using a computer to search for numerical instability in algebraic processes, and so to an extent they are concerned with "automatic rounding error analysis." In [37] Miller defines a quantity $\sigma(d)$ that bounds, to first order, the sensitivity of an algorithm to perturbations in the data d and in the intermediate quantities that the algorithm generates. He then defines the forward stability measure $\rho(d) = \sigma(d)/\kappa(d)$, where $\kappa(d)$ is a condition number for the problem under consideration. The algorithms to be analyzed are required to contain no loops or conditional branches and are presented to Miller's Fortran software in a special numeric encoding. The software automatically computes the partial derivatives needed to evaluate $\rho(d)$, and attempts to maximize ρ using the method of alternating directions. Miller gives several examples illustrating the scope of his software; he shows, for example, that it can identify the instability of the Gram–Schmidt method for orthogonalizing a set of vectors.

In [38] Miller and Spooner extend the work in [37] in several ways. The algorithm to be analyzed is expressed in a Fortran-like language that allows for-loops but not logical tests. The definition of ρ is generalized, and a method of computing it is developed that involves solving a generalized eigenvalue problem. The book [39] gives a thorough development of the work of [38] and provides further examples of the use of the software. The potential of Miller and Spooner's software for exposing numerical instability is clearly demonstrated in [37], [38], and [39], yet the software has apparently not been widely used. We suspect this is largely due to the inability of the software to analyze algorithms expressed in Fortran, or any other standard language.

A different approach to algorithm analysis is taken in [35], [36]. Here errors are measured in a relative rather than an absolute sense, and the stability is analyzed at *fixed data* instead of attempting to maximize instability over all data; however, the analysis is still linearized.

Statistical modelling of rounding errors in an algorithm has been developed by Chatelin and Brunet, and by Vignes. Their techniques involve randomly perturbing the result of every floating point operation and using statistics to measure the effect on the output of the algorithm; see [9], [10], and the references therein. In [8] Fortran preprocessor tools are developed for implementing the statistical approach described in [9] and the local relative error approach of [36].

We take an approach different from those described above. We note that for many algorithms one can define an *easily computable* function f that gives an a posteriori measure of the degree of success or the stability of the algorithm. Our approach is to try to maximize f over the space of problem data using direct search and any available implementation of the algorithm. This approach has several advantages.

- Any algorithm for which a suitable f can be defined and computed can be tested. There are no constraints on the algorithm or the choice of f.

- When f measures numerical stability the interpretation of the results is straightforward, since f reflects the actual rounding errors sustained, instead of being a bound from a linearized or statistical model of rounding error propagation.

- Existing software implementing the algorithm can be utilized.

In a recent Ph.D. thesis Rowan [47] develops another way to search for numerical instability. For an algorithm with data d he maximizes $S(d) = e(d)/\kappa(d)$ using a new direct search maximizer called the subplex method (which is based on the Nelder–Mead simplex method [41]). Here, $e(d) = y_{\text{acc}} - \widehat{y}$ is an approximation to the forward error in the computed solution $\widehat{y}$, where y_{acc} is a more accurate estimate of the true solution than $\widehat{y}$, and the condition number $\kappa(d)$ is estimated using finite difference approximations. The quantity $S(d)$ is a lower bound on the backward error of the algorithm at d. Fortran software given in [47] implements this "functional stability analysis." The software takes as input two user-supplied Fortran subprograms; one implements the algorithm to be tested in single precision, and the other provides a more accurate solution, typically by executing the same algorithm in double precision. The examples in [47] show that Rowan's software is capable of detecting numerical instability in a wide variety of numerical algorithms. Rowan also gives two specific examples of the approach we are advocating here: he uses the subplex method to find a matrix for which the LINPACK condition estimator performs poorly (see §4), and to find unit upper triangular matrices R with $|r_{ij}| \le 1$ that maximize $\kappa_1(R)$.

3. Two direct search methods. We have experimented with two direct search methods. The first is the alternating directions (AD) method. Given a starting value x it attempts to solve the problem (1.1) by repeatedly maximizing over each coordinate direction in turn:

```
repeat
    % One iteration comprises a loop over all components of $x$.
    for $i = 1\colon n$
        find $\alpha$ such that $f(x + \alpha e_i)$ is maximized (line search)
        set $x \leftarrow x + \alpha e_i$
    end
until converged
```

AD is one of the simplest of all optimization methods and its fundamental weakness, that it ignores any interactions between the variables, is well known. Despite the poor reputation of AD we have found that it can perform well on the types of problems considered here. In our implementation of AD the line search is done using a crude scheme that begins by evaluating $f(x + he_i)$ with $h = 10^{-4}x_i$ (or $h = 10^{-4}\max(\|x\|_\infty, 1)$ if $x_i = 0$); if $f(x + he_i) \le f(x)$ then the sign of h is reversed. Then if $f(x + he_i) > f(x)$, h is doubled at most 25 times until no further increase in f is obtained. Our convergence test checks for a sufficient relative increase in f between one iteration and the next: convergence is declared when

$$f_k - f_{k-1} \le \text{tol}\,|f_{k-1}|, \tag{3.1}$$

where f_k is the highest function value at the end of the kth iteration. The AD method has the very modest storage requirement of just a single n-vector.

The second method is the multidirectional search method (MDS) of Dennis and Torczon [55], [56]. This method employs a simplex, which is defined by $n + 1$ vectors $\{v_i\}_0^n$ in $\mathbb{R}^n$. One iteration in the case $n = 2$ is represented pictorially in Fig. 3.1, and may be explained as follows.

The initial simplex is $\{v_0, v_1, v_2\}$ and it is assumed that $f(v_0) = \max_i f(v_i)$. The purpose of an iteration is to produce a new simplex at one of whose vertices f exceeds $f(v_0)$. In the first step the vertices v_1 and v_2 are reflected about v_0 along the lines

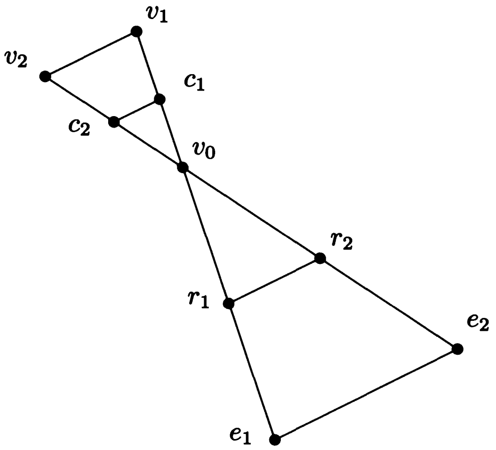


FIG. 3.1. *The possible steps in one iteration of the* MDS *method when* $n = 2$.

joining them to v_0, yielding r_1 and r_2 and the reflected simplex $\{v_0, r_1, r_2\}$. If this reflection step is successful, that is, if $\max_i f(r_i) > f(v_0)$, then the edges from v_0 to r_i are doubled in length to give an expanded simplex $\{v_0, e_1, e_2\}$. The original simplex is then replaced by $\{v_0, e_1, e_2\}$ if $\max_i f(e_i) > \max_i f(r_i)$, and otherwise by $\{v_0, r_1, r_2\}$. If the reflection step is unsuccessful then the edges $v_0 - v_i$ of the original simplex are shrunk to half their length to give the contracted simplex $\{v_0, c_1, c_2\}$. This becomes the new simplex if $\max_i f(c_i) > \max_i f(v_i)$, in which case the current iteration is complete; otherwise the algorithm jumps back to the reflection step, now working with the contracted simplex. For further details of the MDS method, see [15], [55], and [56].

The MDS method requires at least $2n$ independent function evaluations per iteration, which makes it very suitable for parallel implementation. Generalizations of the MDS method that are even more suitable for parallel computation are described in [15]. The MDS method requires $O(n^2)$ elements of storage for the simplices, but this can be reduced to $O(n)$ (at the cost of extra bookkeeping) if an appropriate choice of initial simplex is made [15].

Unusually for a direct search method, the MDS method possesses some convergence theory. Torczon [56] shows that if the level set of f at v_0^0 is compact and f is continuously differentiable on this level set then a subsequence of the points v_0^k (where k denotes the iteration index) converges to a stationary point of f. Moreover, she gives an extension of this result that requires only continuity of f and guarantees convergence to either a stationary point of f or a point where f is not continuously differentiable. No such convergence results are known for the Nelder–Mead direct search method [16], [41], which also employs a simplex but which is fundamentally different from the MDS method. Our limited experiments with the Nelder–Mead method indicate that while it can sometimes outperform the MDS method, the MDS method is generally superior for our purposes.

Our implementation of the MDS method provides two possible starting simplices, both of which include the starting point x_0: a regular one (all sides of equal length) and a right-angled one based on the coordinate axes, both as described in [55]. The scaling is such that each edge of the regular simplex, or each edge of the right-angled simplex that is joined to x_0, has length $\max(\|x_0\|_\infty, 1)$. Also as in [55], the main termination test halts the computation when the relative size of the simplex is no larger than a tolerance tol, that is, when

$$\frac{1}{\max(1, \|v_0\|_1)} \max_{1 \le i \le n} \|v_i - v_0\|_1 \le \text{tol}. \tag{3.2}$$

Unless otherwise stated, we used tol $= 10^{-3}$ in (3.1) and (3.2) in all our experiments.

It is interesting to note that the MDS method and our particular implementation of the AD method do not exploit the numerical values of f: their only use of f is to compare two function values to see which is the larger!

Our MATLAB implementations of the AD and MDS methods can be obtained from netlib [18] by sending electronic mail to `netlib@ornl.gov` comprising the message **`send dsmax from matlab/optimization`**.

4. Condition estimators. Condition estimation is the problem of computing an inexpensive but "reliable" estimate of $\kappa(A) = \|A\|\|A^{-1}\|$, for some matrix norm, given a factorization of the nonsingular matrix A. (Other condition numbers of A are also of interest, but we will concentrate on this standard condition number.) The best known condition estimator is the one used in LINPACK [17]; it makes use of an LU factorization of A and works with the 1-norm. Its development is described in [12].[3] Several years after [12] was published several counterexamples to the LINPACK condition estimator were discovered by Cline and Rew [13]; by a counterexample we mean a parametrized matrix for which the quotient "condition estimate divided by true condition number" can be made arbitrarily small (or large, depending on whether the estimator produces a lower bound or an upper bound) by varying a parameter. Despite the existence of these counterexamples the LINPACK estimator has been widely used and is regarded as being almost certain to produce an estimate correct to within a factor ten in practice [27].

Another 1-norm condition estimation algorithm was developed by Higham [29], [30], building on an algorithm of Hager [25]. This estimator is in the NAG library and is being used throughout LAPACK [2]. The general algorithm estimates $\|B\|_1$ given a means for forming matrix-vector products Bx and B^Ty. By taking $B = A^{-1}$ and using an LU factorization of A we obtain an estimator with the same functionality as the LINPACK estimator. Counterexamples to the general algorithm are identified in [29].

A 2-norm condition estimator was developed by Cline, Conn, and Van Loan [11, Algorithm 1]; see also [58]. The algorithm builds on the ideas underlying the LINPACK estimator and estimates $\sigma_{\min}(R) = \|R^{-1}\|_2^{-1}$ or $\sigma_{\max}(R) = \|R\|_2$ for a triangular matrix R. Here, $\sigma_{\min}$ and $\sigma_{\max}$ denote the smallest and largest singular values, respectively. Full matrices can be treated if a factorization $A = QR$ is available (Q orthogonal, R upper triangular), since R and A have the same singular values. The estimator performs extremely well in numerical tests [11], [27], often producing an estimate having some correct digits. No counterexamples to the estimator were known until Bischof [5] obtained counterexamples as a by-product of the analysis of a different, but related, method.

We have experimented with MATLAB implementations of the three condition estimators discussed above. RCOND is the LINPACK estimator as built into MATLAB. SONEST implements the algorithm of [29] as applied to estimating $\kappa_1(A)$. SIGMAN

[3] It is not widely known that a precursor to the LINPACK condition estimator is presented in [24]. I thank G. W. Stewart for pointing this out to me.

is an implementation of the algorithm of [11] for estimating $\sigma_{\min}(R)$, where R is upper triangular.

For RCOND we define $x = \operatorname{vec}(A)$, $A \in \mathbb{R}^{n\times n}$, and

$$f(x) = \frac{\kappa_1(A)}{\operatorname{est}(A)},$$

where $\operatorname{est}(A) \le \kappa_1(A)$ is the condition estimate computed by RCOND. The same definition is used for SONEST, which also computes a lower bound. We note that since the algorithms underlying RCOND and SONEST contain tests and branches, for certain A an arbitrarily small change in A can completely change the condition estimate; hence for both algorithms f has points of discontinuity.

Since SIGMAN was designed for upper triangular matrices $R \in \mathbb{R}^{n\times n}$ we take in this case $x = \overline{\operatorname{vec}}(R)$, where $\overline{\operatorname{vec}}$ is the vec operator modified to skip over elements in the lower triangle of its argument, and we define

$$f(x) = \frac{\operatorname{est}(R)}{\sigma_{\min}(R)},$$

where $\operatorname{est}(R) \ge \sigma_{\min}(R)$ is the estimate.

We applied the MDS maximizer to RCOND starting at $A = I_4$. After 30 iterations and 1009 function evaluations the maximizer had located the matrix

$$A = \begin{bmatrix} 1.1380 & 0.1380 & -2.52\times 10^{6} & 0.1380 \\ 0.1380 & 1.1380 & -1.34\times 10^{7} & 0.1380 \\ 0.1380 & 0.1380 & 1.1380 & 0.1380 \\ 0.1380 & 0.1380 & 1.59\times 10^{7} & 1.1380 \end{bmatrix},$$

for which

$$\kappa_1(A) = 9.88\times 10^{14}, \quad \operatorname{est}(A) = 2.17\times 10^{10}, \quad \frac{\kappa_1(A)}{\operatorname{est}(A)} = 4.56\times 10^{4}.$$

(For comparison, with the same starting matrix the AD maximizer yielded $f = 18.2$ after six iterations and 950 function evaluations.) This matrix A is badly scaled, and its validity as a counterexample could be questioned on the grounds that for linear equations the condition number $\kappa_1(A)$ is not necessarily an appropriate measure of problem sensitivity when A is badly scaled (see, for example, [22, §3.5.2]). This objection can be overcome by maximizing f subject to a constraint that ensures A is reasonably well scaled. A simple, but effective, constraint is $\kappa_1(A) \le \theta$, where θ is a suitable tolerance. To incorporate this constraint we use a crude penalty function approach in which f is redefined so that $f(x) = -10^{300}$ whenever the constraint is violated. Applying the MDS maximizer with starting matrix $A = \operatorname{diag}(1,-1,1,-1)$, $\theta = 10^4$, and $\mathrm{tol} = 10^{-9}$ in (3.2), we obtained after 124 iterations and 4625 function evaluations the well-scaled matrix

$$A = \begin{bmatrix} 0.1094 & 0.4559 & 0.1434 & 0.1461 \\ 1.1989 & -0.8617 & 0.1359 & 0.1383 \\ 0.1375 & 2.2531 & 2.2017 & 0.1383 \\ 0.1404 & -2.6932 & 0.1383 & -0.8617 \end{bmatrix},$$

for which

$$\kappa_1(A) = 7.47\times 10^{3}, \quad \operatorname{est}(A) = 5.37, \quad \frac{\kappa_1(A)}{\operatorname{est}(A)} = 1.39\times 10^{3}.$$

We note that the parametrized counterexamples in [13] all become badly scaled when the parameter is chosen to make the condition estimate poor. The only previously known well-scaled counterexample to the LINPACK condition estimator is an $n \times n$ lower triangular matrix L in [13] for which $\kappa_1(L)/\text{est}(L) = 2^{n-1}$.

For SONEST, the two maximizers make extremely slow progress starting with $A = I_4$. A better starting value for both maximizers is the 4×4 version of the $n \times n$ matrix with $a_{ij} = \cos((i-1)(j-1)\pi/(n-1))$ [34]. After 11 iterations and 1001 function evaluations the AD maximizer had determined a (well-scaled) matrix A for which

$$\kappa_1(A) = 2.94 \times 10^5, \quad \text{est}(A) = 4.81, \quad \frac{\kappa_1(A)}{\text{est}(A)} = 6.11 \times 10^4.$$

Applying MDS to SIGMAN, starting with $R = I_4$, we obtained after 65 iterations and 1511 function evaluations a matrix R such that

$$\sigma_{\min}(R) = 3.25 \times 10^{-1}, \quad \text{est}(R) = 2.00 \times 10^1, \quad \frac{\text{est}(R)}{\sigma_{\min}(R)} = 6.16 \times 10^1.$$

Using this matrix to start the AD maximizer led after two iterations and a further 93 function evaluations to $\overline{R}$ such that

$$\sigma_{\min}(\overline{R}) = 3.25 \times 10^{-1}, \quad \text{est}(\overline{R}) = 4.31 \times 10^1, \quad \frac{\text{est}(\overline{R})}{\sigma_{\min}(\overline{R})} = 1.33 \times 10^2.$$

These results are surprising. With little effort on our part in the choice of starting matrix the maximizers have discovered examples where each of the condition estimators fails to achieve its objective of producing an estimate correct to within an order of magnitude. Such numerical examples have apparently never been observed in practical computation, or in tests with random matrices such as those in [27]. The value of direct search maximization in this context is clear: it can readily demonstrate the fallibility of a condition estimator—a task that can be extremely difficult to accomplish using theoretical analysis or tests with random matrices. Moreover, the numerical examples obtained from direct search may provide a starting point for the construction of parametrized theoretical ones, or for the improvement of a condition estimation algorithm.

An area of current research in condition estimation is the derivation of algorithms appropriate in applications such as signal processing where a matrix undergoes repeated low rank updates. Several algorithms have been developed [42], [50], but counterexamples to them are not known. Direct search on the appropriate ratio f could provide further insight into these methods.

We note that the direct search approach provides an alternative to the usual way of assessing the quality of condition estimators, which is to examine the quality of the estimates produced for random matrices [27]. One could instead measure the difficulty that an optimizer has in "defeating" a condition estimator—perhaps over a large number of trials with random starting matrices.

As well as measuring the quality of a single algorithm, direct search can be used to compare two competing algorithms, in order to investigate whether one algorithm performs uniformly better than the other. We applied the MDS maximizer to the function

$$f(x) = \frac{\text{estS}(A)}{\text{estR}(A)},$$

where estS(A) and estR(A) are the condition estimates from SONEST and RCOND, respectively. If $f(x) > 1$ then SONEST has produced a larger lower bound for $\kappa_1(A)$ than RCOND. Starting with $A = I_4$, the MDS maximizer converged after 22 iterations to a matrix A for which $\mathrm{estS}(A) = \kappa_1(A)$ and $f(x) = 150.1$. With f defined as $f(x) = \mathrm{estR}(A)/\mathrm{estS}(A)$, and starting with a random matrix $B \in \mathbb{R}^{4\times 4}$, the MDS maximizer converged after 40 iterations to a matrix with $f(x) = 63.90$. This experiment shows that neither estimator is uniformly superior to the other. This conclusion would be onerous to reach by theoretical analysis of the algorithms.

Finally, we use direct search to investigate an open question raised in [27]. If $A\Pi = QR$ is a QR factorization with column pivoting of $A \in \mathbb{R}^{n\times n}$, then [27, Thm. 6.2]

$$\frac{1}{|r_{nn}|} \le \|R^{-1}\|_{1,2} \le \frac{2^{n-1}}{|r_{nn}|}. \tag{4.1}$$

Moreover, there is much experimental evidence to show that $1/|r_{nn}|$ is rarely more than ten times smaller than $\|R^{-1}\|_2$. The question raised in [27] is whether for R from the QR factorization with column pivoting the estimate $\mathrm{est}(R) \le \kappa_1(R)$ produced by the LINPACK estimator has the desirable property that

$$\mathrm{est}(R) \ge \frac{\|R\|_1}{|r_{nn}|}, \tag{4.2}$$

that is, whether the LINPACK estimator always performs at least as well as the trivial lower bound from (4.1). We applied the MDS maximizer to $f(x) = |r_{nn}|^{-1}\|R\|_1/\mathrm{est}(R)$, where $A \in \mathbb{R}^{n\times n}$, $x = \mathrm{vec}(A)$, and $A\Pi = QR$. Starting with $A = I_4$, the maximizer achieved $f(x) = 520.4$ after 67 iterations and 2929 function evaluations. Thus (4.2) is not satisfied—not even to within a reasonable constant factor. However, we have not been able to generate any matrix A for which SIGMAN produces an estimate $\mathrm{est}(R) > |r_{nn}|$ (with $A\Pi = QR$), so it is an open question as to whether $\mathrm{est}(R) \le |r_{nn}|$ always holds for SIGMAN.

5. Other topics. In this section we describe five further topics in matrix computations in which direct search yields interesting results. The first four examples are all concerned with the instability of an algorithm in the presence of rounding errors; here, unlike in the applications considered so far, the objective function depends in an essential way on rounding errors. In our MATLAB computing environment the unit roundoff $u \approx 1.11 \times 10^{-16}$.

5.1. Fast matrix inversion. First, we discuss an example for which there is no existing error analysis, and for which direct search reveals numerical instability. In [51] Strassen gives a method for multiplying two $n\times n$ matrices in $O(n^{\log_2 7})$ operations ($\log_2 7 \approx 2.801$); he also gives a method for inverting an $n \times n$ matrix with the same asymptotic cost. The inversion method is based on the following formulae, where

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad A_{ij} \in \mathbb{R}^{m\times m}, \quad n = 2m,$$

and $C = A^{-1}$:

$$P_1 = A_{11}^{-1}, \quad P_2 = A_{21}P_1,$$
$$P_3 = P_1A_{12}, \quad P_4 = A_{21}P_3,$$

$$P_5 = P_4 - A_{22}, \quad P_6 = P_5^{-1},$$

$$C = \begin{bmatrix} P_1 - P_3P_6P_2 & P_3P_6 \\ P_6P_2 & -P_6 \end{bmatrix}.$$

(These formulae are easily derived via a block *LU* factorization of A.) The matrix multiplications are done by Strassen's method and the inversions determining P_1 and P_6 are done by recursive invocations of the method itself. The inversion method is clearly unstable for general A, because the method breaks down if A_{11} is singular. Indeed Strassen's inversion method has been implemented on a Cray-2 in [3] and tested for $n \le 2048$, and it is observed empirically in [3] that the method has poor numerical stability. Here we use direct search to investigate the numerical stability.

With $x = \mathrm{vec}(A) \in \mathbb{R}^{n^2}$, define the stability measure

$$f(x) = \frac{\min\{\|A\widehat{C} - I\|_\infty, \|\widehat{C}A - I\|_\infty\}}{\|A\|_\infty \|\widehat{C}\|_\infty}, \tag{5.1}$$

where $\widehat{C}$ is the inverse of A computed using Strassen's inversion method. This definition of f is appropriate because, as shown in [19], for most conventional matrix inversion methods either the left residual $\widehat{C}A - I$ or the right residual $A\widehat{C} - I$ is guaranteed to have norm of order $u\|\widehat{C}\|\|A\|$. To treat Strassen's inversion method as favourably as possible we use just one level of recursion; thus P_1 and P_6 are computed using Gaussian elimination with partial pivoting, but the multiplications are done with Strassen's method. We applied the MDS maximizer, with tol $= 10^{-9}$ in (3.2), starting with the 4×4 Vandermonde matrix whose (i, j) element is $((j-1)/3)^{i-1}$. After 34 iterations the maximizer had converged with $f = 0.838$, which represents complete instability. The corresponding matrix A is well conditioned, with $\kappa_2(A) = 82.4$. For comparison, the value of f when A is inverted using Strassen's method with *conventional* multiplication is $f = 6.90 \times 10^{-2}$; this confirms that the instability is not due to the use of fast multiplication techniques—it is inherent in the inversion formulae.

If A is a symmetric positive definite matrix then its leading principal submatrices are no more ill conditioned than the matrix itself, so one might expect Strassen's inversion method to be stable for such matrices. To investigate this possibility we carried out the same maximization as before except we enforced positive definiteness as follows: when the maximizer generates a vector $x \equiv \mathrm{vec}(B)$, A in (5.1) is defined as $A = B^TB$. Starting with a 4×4 random matrix A with $\kappa_2(A) = 6.71 \times 10^7$ the maximization yielded the value $f = 3.32 \times 10^{-8}$ after 15 iterations, and the corresponding value of f when conventional multiplication is used is $f = 6.61 \times 10^{-11}$ (the "maximizing" matrix A has condition number $\kappa_2(A) = 3.58 \times 10^9$).

The conclusion from these experiments is that Strassen's inversion method cannot be guaranteed to produce a small left or right residual even when A is symmetric positive definite and conventional multiplication is used. Hence the method must be regarded as being fundamentally unstable. (The analyses in [14, §2.3] and [19, §2.2] can be used to obtain further insight into this instability.)

5.2. Fast Vandermonde system solvers.

In 1970 Björck and Pereyra [7] published two algorithms for solving the Vandermonde systems $Vx = b$ and $V^Ta = f$ in $O(n^2)$ operations, where $V = (\alpha_j^{i-1}) \in \mathbb{R}^{n \times n}$. These algorithms have been used in various applications [1], [48], [57] and generalized in several ways [6], [28], [31], [54], and their numerical stability has been investigated [26], [31], [59]. In [7] it was pointed out that the algorithms sometimes produce surprisingly accurate results, and an explanation for this was given in [26]. However, analysis in [31] predicts that the algorithms can be moderately unstable, and an example of instability is given in [31], together with suggested remedies. The example of [31] appears to be rare since no instances of instability of the Björck–Pereyra algorithms were reported in the first twenty years following their publication. It is therefore interesting to see whether instability can be located by direct search.

Consider the dual system $V^T a = f$ and define the relative residual

$$\begin{aligned} g(x) &\equiv \frac{\|V^T\widehat{a} - f\|_\infty}{\|V^T\|_\infty \|\widehat{a}\|_\infty + \|f\|_\infty} \\ &= \min\{\epsilon : (V + \Delta V)^T \widehat{a} = f + \Delta f, \\ &\qquad \|\Delta V\|_\infty \le \epsilon \|V\|_\infty,\ \|\Delta f\|_\infty \le \epsilon \|f\|_\infty\}, \end{aligned}$$

where $x = [\alpha_1, \dots, \alpha_n, f_1, \dots, f_n]^T \in \mathbb{R}^{2n}$ and $\widehat{a}$ is the computed solution from the Björck–Pereyra dual algorithm. The equality above says that the relative residual is equal to the normwise backward error and is well known (see [45]). It is desirable to constrain the points α_i to be distinct and in increasing order (since this ordering is standard and usually helps the numerical stability [31]). To do so we redefine g so that $g(x) = -10^{300}$ if the ordering conditions are not satisfied. We applied direct search to g with $n = 25$, starting with the points α_i equally spaced on $[-1, 1]$ and with $f_i \equiv 1$. Using a combination of MDS and AD maximizer invocations we obtained the value $g(x) = 2.57 \times 10^{-13}$, after a total of approximately 2100 function evaluations.

Thus, given "innocuous" starting values, the maximizers find moderate instability of size three orders of magnitude. The vector produced by the maximization does not represent a pathological problem: the points α_i are approximately equally spaced between -0.996 and 0.588, the vector f has positive elements lying between 0.309 and 2.42, and the Vandermonde matrix V satisfies $\kappa_2(V) = 2.27 \times 10^{10}$. The $n = 25$ problem specified in [31, eq. (6.2)] yields $g(x) = 2.03 \times 10^{-12}$, so the value of g located by direct search is not a global maximum, but it is sufficiently large to reveal instability. We also tried using the problem just mentioned as a starting point for direct search. The AD maximizer increased g to 6.79×10^{-12} in two iterations, but was unable to increase g further.

5.3. Matrix inverse. Numerical analysts universally deprecate the idea of solving a linear system $Ax = b$ by forming $x = A^{-1} \times b$. The reasons are that Gaussian elimination with partial pivoting (GEPP) is generally less expensive and more numerically stable than use of a computed matrix inverse. The difference in computational expense is easily explained and demonstrated (classically, the difference is a factor of 3). The difference in numerical stability is more subtle, and is rarely discussed in the literature, although an excellent analysis is given in [21, §4.7.2]. The instability of the inversion approach is easily demonstrated using direct search. For the linear system $Ay = b$ where $A \in \mathbb{R}^{n \times n}$ let

$$f(x) = \frac{\|b - A\widehat{y}\|_\infty}{\|A\|_\infty \|\widehat{y}\|_\infty + \|b\|_\infty},$$

where A^{-1} is computed via GEPP, $\widehat{y}$ is the computed version of $y = A^{-1} \times b$, and $x \in \mathbb{R}^{n^2+n}$ contains the elements of A and b. We applied the MDS maximizer, with $\text{tol} = 10^{-9}$ in (3.2), taking as initial data the Hilbert matrix of order 4 and the vector of all ones. After 27 iterations and 1741 function evaluations the maximizer converged with $f(x) = 2.91 \times 10^{-9}$. In contrast, with y computed from GEPP, and using the same starting values, the maximizer was unable to drive f above the unit roundoff level 1.11×10^{-16}.

For the matrix inversion method the final A and b found by the maximizer satisfy

$$\kappa_2(A) = 1.05 \times 10^{10}, \quad \|A\|_2\|y\|_2 = 40.2, \quad \|b\|_2 = 2.50;$$

thus b is a right-hand side for which y does not reflect the ill condition of A. As explained in [21], it is for such A and b that the instability of matrix inversion as a means of solving $Ax = b$ is most pronounced.

5.4. Complex matrix multiplication. It is well known that two complex numbers can be multiplied using only three real multiplications. An analogous result holds for matrices. Let $A = A_1 + iA_2, B = B_1 + iB_2$, where $A_i, B_i \in \mathbb{R}^{n\times n}$. If $C = C_1 + iC_2 = AB$, then

$$C_1 = A_1B_1 - A_2B_2, \tag{5.2a}$$

$$C_2 = (A_1 + A_2)(B_1 + B_2) - A_1B_1 - A_2B_2, \tag{5.2b}$$

and these expressions can be evaluated using three real matrix multiplications and five real matrix additions. This yields a saving in arithmetic operations of about 25 percent compared to the usual way of forming C, which involves four real matrix multiplications. However, this alternative method is known to be less stable than conventional multiplication in the sense that the computed imaginary part of the product can be relatively inaccurate [33]. Let

$$f(x) = \frac{\|\widehat{C}_2 - \overline{C}_2\|_\infty}{\|\widehat{C}_2\|_\infty},$$

where $x = [\text{vec}(A_1)^T \ \text{vec}(A_2)^T \ \text{vec}(B_1)^T \ \text{vec}(B_2)^T]^T \in \mathbb{R}^{4n^2}$ and $\widehat{C}_2$ and $\overline{C}_2$ are the computed imaginary parts from conventional multiplication and the formula (5.2b), respectively. A large value for $f(x)$ implies that $\overline{C}_2$ is inaccurate, since we know from standard error analysis that $\widehat{C}_2$ will always be as accurate as can be expected. To test how readily direct search can expose the instability of (5.2a) we applied the MDS maximizer to f with the starting data $A = E + i\alpha E$, $B = E + i\alpha E$, where $E \in \mathbb{R}^{n\times n}$ is the matrix of 1's. With $n = 4$, tol $= 10^{-6}$ in (3.2), and $\alpha = 1$, the maximizer converged after 12 iterations with $f(x) = 3.40 \times 10^{-16}$. With the same n and tol, and $\alpha = 5$, the maximizer converged with $f(x) = 9.99 \times 10^{-11}$ after 24 iterations. Thus, with a little experimentation in the choice of starting value, the instability is easily revealed.

5.5. QR factorization. If $A = QR \in \mathbb{R}^{n\times n}$ is a QR factorization then $\sigma_{\min}(A) = \sigma_{\min}(R) \le \min_i |r_{ii}|$. If column pivoting is used in the QR factorization then this inequality differs from equality by at most a factor 2^{n-1}, as shown by (4.1). But in general the inequality can be arbitrarily weak, as is well known. This is easily confirmed by direct search. Let $f(x) \equiv \min_i |r_{ii}|/\sigma_{\min}(A)$, where $x = \text{vec}(A)$ and $A = QR$. We applied the MDS maximizer followed by the AD maximizer, with tol $= 10^{-3}$ and starting with $A = I_4$, and we obtained $f(x) = 6.58 \times 10^7$ after approximately 930 function evaluations. In fact, the maximizers make rapid progress in increasing f for every starting value we have tried. This is perhaps not surprising, since Foster

[20] gives bounds on the probability that f exceeds β for a class of random matrices, and the probabilities are significant even for large β. For the QR factorization with column pivoting, starting with $A = I_4$, the same maximization produced $f(x) = 2.57$, which is well short of the global maximum $2^{n-1} = 8$.

6. Conclusions. Our experience in using direct search methods has convinced us that they are a useful supplement to the traditional means of analyzing algorithms in matrix computations, such as rounding error analysis and numerical testing with random data. Indeed, tests with random data tend to reveal the average-case behaviour of an algorithm, but the worst-case is also of interest. An underlying theme of this work is that direct search can be vastly more effective than Monte Carlo testing at revealing worst-case behaviour (this is particularly true for the condition estimators of §3).

As we have shown, direct search is sometimes capable of exposing failure of algorithms even when given a trivial starting value such as an identity matrix. An informed choice of starting value coming from partial understanding of the algorithm increases the chance of a revealing optimization. Unsuccessful optimizations can also provide useful information. As Miller and Spooner explain [38, p. 370], "Failure of the maximizer to find large values of ω (say) can be interpreted as providing evidence for stability equivalent to a large amount of practical experience with low-order matrices."

To make use of direct search one has to be able to express the question of interest as an unconstrained optimization problem. As we have shown, this can often be done by employing an appropriate residual or overestimation ratio. If numerical stability is of interest and a suitable objective function cannot be defined then the approach of Rowan [47] is attractive, since it automatically constructs stability estimates given only the ability to execute the algorithm at two different precisions.

Direct search optimization is potentially useful in other areas of numerical analysis besides matrix computations. An experiment in which direct search is used to reveal the fallibility of an adaptive quadrature routine is described in [47]; direct search is used to investigate the accuracy of floating point summation in [32]; and direct search has been used to help tune heuristic parameters in Fortran codes for the numerical solution of ordinary differential equations [49]. We hope that in addition to encouraging researchers in numerical analysis to experiment with direct search optimization, this work will encourage optimization researchers to devote more attention to the rather neglected area of direct search. The multidirectional search method of Dennis and Torczon performed extremely well in our experiments, and alternating directions performed much better than the textbooks might lead one to expect. Parallel direct search methods, such as those in [15], seem particularly attractive for tackling difficult problems such as maximizing the growth factor for Gaussian elimination with complete pivoting [23].

Acknowledgments. I thank Des Higham and Nick Trefethen for their many helpful comments on this work.

REFERENCES

[1] M. Almacany, C. B. Dunham, and J. Williams, *Discrete Chebyshev approximation by interpolating rationals*, IMA J. Numer. Anal., 4 (1984), pp. 467–477.

[2] E. ANDERSON, Z. BAI, C. H. BISCHOF, J. W. DEMMEL, J. J. DONGARRA, J. J. DU CROZ, A. GREENBAUM, S. J. HAMMARLING, A. MCKENNEY, S. OSTROUCHOV, AND D. C. SORENSEN, *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[3] D. H. BAILEY AND H. R. P. FERGUSON, *A Strassen–Newton algorithm for high-speed parallelizable matrix inversion*, in Proceedings of Supercomputing '88, IEEE Computer Society Press, New York, 1988, pp. 419–424.

[4] M. C. BERENBAUM, *Direct search methods in the optimisation of cancer chemotherapy*, Br. J. Cancer, 61 (1991), pp. 101–109.

[5] C. H. BISCHOF, *Incremental condition estimation*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 312–322.

[6] A. BJÖRCK AND T. ELFVING, *Algorithms for confluent Vandermonde systems*, Numer. Math., 21 (1973), pp. 130–137.

[7] A. BJÖRCK AND V. PEREYRA, *Solution of Vandermonde systems of equations*, Math. Comp., 24 (1970), pp. 893–903.

[8] B. BLISS, M.-C. BRUNET, AND E. GALLOPOULOS, *Automatic program instrumentation with applications in performance and error analysis*, in Expert Systems for Scientific Computing, E. N. Houstis, J. R. Rice, and R. Vichnevetsky, eds., North-Holland, 1992, pp. 235–260.

[9] M.-C. BRUNET, *Contribution à la Fiabilité de Logiciels Numériques et à L'analyse de Leur Comportement: Une Approche Statistique*, Ph.D. thesis, U.E.R. Mathematiques de la decision, Université de Paris IX Dauphine, Jan. 1989.

[10] F. CHATELIN AND M.-C. BRUNET, *A probabilistic round-off error propagation model. Application to the eigenvalue problem*, in Reliable Numerical Computation, M. G. Cox and S. J. Hammarling, eds., Oxford University Press, Oxford, U.K., 1990, pp. 139–160.

[11] A. K. CLINE, A. R. CONN, AND C. F. VAN LOAN, *Generalizing the LINPACK condition estimator*, in Numerical Analysis, Mexico 1981, J. P. Hennart, ed., Lecture Notes in Mathematics 909, Springer-Verlag, Berlin, 1982, pp. 73–83.

[12] A. K. CLINE, C. B. MOLER, G. W. STEWART, AND J. H. WILKINSON, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368–375.

[13] A. K. CLINE AND R. K. REW, *A set of counter-examples to three condition number estimators*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 602–611.

[14] J. W. DEMMEL AND N. J. HIGHAM, *Stability of block algorithms with fast level* 3 *BLAS*, ACM Trans. Math. Software, 18 (1992), pp. 274–291.

[15] J. E. DENNIS, JR. AND V. J. TORCZON, *Direct search methods on parallel machines*, SIAM J. Optimization, 1 (1991), pp. 448–474.

[16] J. E. DENNIS, JR. AND D. J. WOODS, *Optimization on microcomputers: The Nelder–Mead simplex algorithm*, in New Computing Environments: Microcomputers in Large-Scale Computing, A. Wouk, ed., Society for Industrial and Applied Mathematics, Philadelphia, 1987, pp. 116–122.

[17] J. J. DONGARRA, J. R. BUNCH, C. B. MOLER, AND G. W. STEWART, *LINPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1979.

[18] J. J. DONGARRA AND E. GROSSE, *Distribution of mathematical software via electronic mail*, Comm. ACM, 30 (1987), pp. 403–407.

[19] J. J. DUCROZ AND N. J. HIGHAM, *Stability of methods for matrix inversion*, IMA J. Numer. Anal., 12 (1992), pp. 1–19.

[20] L. V. FOSTER, *The probability of large diagonal elements in the* QR *factorization*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 531–544.

[21] P. E. GILL, W. MURRAY, AND M. H. WRIGHT, *Numerical Linear Algebra and Optimization, Volume* 1, Addison-Wesley, Reading, MA, 1991.

[22] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd Ed., Johns Hopkins University Press, Baltimore, MD, 1989.

[23] N. I. M. GOULD, *On growth in Gaussian elimination with complete pivoting*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 354–361.

[24] W. B. GRAGG AND G. W. STEWART, *A stable variant of the secant method for solving nonlinear equations*, SIAM J. Numer. Anal., 13 (1976), pp. 889–903.

[25] W. W. HAGER, *Condition estimates*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 311–316.

[26] N. J. HIGHAM, *Error analysis of the Björck–Pereyra algorithms for solving Vandermonde systems*, Numer. Math., 50 (1987), pp. 613–632.

[27] ———, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596.

[28] N. J. HIGHAM, *Fast solution of Vandermonde-like systems involving orthogonal polynomials*, IMA J. Numer. Anal., 8 (1988), pp. 473–486.

[29] ———, *FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation (Algorithm* 674), ACM Trans. Math. Software, 14 (1988), pp. 381–396.

[30] N. J. HIGHAM, *Experience with a matrix norm estimator*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 804–809.

[31] ———, *Stability analysis of algorithms for solving confluent Vandermonde-like systems*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 23–41.

[32] ———, *The accuracy of floating point summation*, Numerical Analysis Report No. 198, Univ. of Manchester, England, Apr. 1991; SIAM J. Sci. Comput., 14 (1993), to appear.

[33] ———, *Stability of a method for multiplying complex matrices with three real matrix multiplications*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 681–687.

[34] N. J. HIGHAM AND D. J. HIGHAM, *Large growth factors in Gaussian elimination with pivoting,* SIAM J. Matrix Anal. Appl., 10 (1989), pp. 155–164.

[35] J. L. LARSON, M. E. PASTERNAK, AND J. A. WISNIEWSKI, *Algorithm* 594: *Software for relative error analysis*, ACM Trans. Math. Software, 9 (1983), pp. 125–130.

[36] J. L. LARSON AND A. H. SAMEH, *Algorithms for roundoff error analysis—A relative error approach*, Computing, 24 (1980), pp. 275–297.

[37] W. MILLER, *Software for roundoff analysis*, ACM Trans. Math. Software, 1 (1975), pp. 108–128.

[38] W. MILLER AND D. SPOONER, *Software for roundoff analysis,* II, ACM Trans. Math. Software, 4 (1978), pp. 369–387.

[39] W. MILLER AND C. WRATHALL, *Software for Roundoff Analysis of Matrix Algorithms*, Academic Press, New York, 1980.

[40] C. B. MOLER, J. N. LITTLE, AND S. BANGERT, PC-*Matlab User's Guide*, The MathWorks, Inc., Natick, MA, 1987.

[41] J. A. NELDER AND R. MEAD, *A simplex method for function minimization*, Comput. J., 7 (1965), pp. 308–313.

[42] D. J. PIERCE AND R. J. PLEMMONS, *Fast adaptive condition estimation*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 274–291.

[43] M. J. D. POWELL, *A survey of numerical methods for unconstrained optimization*, SIAM Rev., 12 (1970), pp. 79–97.

[44] ———, *A view of unconstrained minimization routines that do not require derivatives*, ACM Trans. Math. Software, 1 (1975), pp. 97–107.

[45] J. L. RIGAL AND J. GACHES, *On the compatibility of a given solution with the data of a linear system*, J. Assoc. Comput. Mach., 14 (1967), pp. 543–548.

[46] M. W. ROUTH, P. A. SWARTZ, AND M. B. DENTON, *Performance of the super modified simplex*, Analytical Chemistry, 49 (1977), pp. 1422–1428.

[47] T. H. ROWAN, *Functional Stability Analysis of Numerical Algorithms*, Ph.D. thesis, Univ. of Texas at Austin, May 1990.

[48] L. F. SHAMPINE, I. GLADWELL, AND R. W. BRANKIN, *Reliable solution of special event location problems for* ODE*s*, ACM Trans. Math. Software, 17 (1991), pp. 11–25.

[49] P. W. SHARP, Personal communcation, 1991.

[50] G. M. SHROFF AND C. H. BISCHOF, *Adaptive condition estimation for rank-one updates of* QR *factorizations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1264–1279.

[51] V. STRASSEN, *Gaussian elimination is not optimal*, Numer. Math., 13 (1969), pp. 354–356.

[52] W. H. SWANN, *Direct search methods*, in Numerical Methods for Unconstrained Optimization, W. Murray, ed., Academic Press, New York, 1972, pp. 13–28.

[53] ———, *Constrained optimization by direct search*, in Numerical Methods for Constrained Optimization, P. E. Gill and W. Murray, eds., Academic Press, New York, 1974, pp. 191–217.

[54] W. P. TANG AND G. H. GOLUB, *The block decomposition of a Vandermonde matrix and its applications*, BIT, 21 (1981), pp. 505–517.

[55] V. J. TORCZON, *Multi-directional Search: A Direct Search Algorithm for Parallel Machines*, Ph.D. thesis, Dept. of Mathematical Sciences, Rice University, Houston, TX, May 1989.

[56] ———, *On the convergence of the multidirectional search algorithm*, SIAM J. Optimization, 1 (1991), pp. 123–145.

[57] G. E. TRAPP AND W. SQUIRE, *Solving nonlinear Vandermonde systems*, Comput. J., 18 (1975), pp. 373–374.

[58] C. F. VAN LOAN, *On estimating the condition of eigenvalues and eigenvectors*, Linear Algebra Appl., 88/89 (1987), pp. 715–732.

[59] J. M. VARAH, *Errors and perturbations in Vandermonde systems*, manuscript, 1990.

[60] J. H. WILKINSON, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 334–352, April 1993

ELIMINATION STRUCTURES FOR UNSYMMETRIC SPARSE *LU* FACTORS[*]

JOHN R. GILBERT[†] AND JOSEPH W. H. LIU[‡]

Abstract. The elimination tree is central to the study of Cholesky factorization of sparse symmetric positive definite matrices. In this paper, the elimination tree is generalized to a structure appropriate for the sparse *LU* factorization of unsymmetric matrices. A pair of directed acyclic graphs, called *elimination dags*, is defined and they are used to characterize the zero-nonzero structures of the lower and upper triangular factors. These elimination structures are applied in a new algorithm to compute fill for sparse *LU* factorization. Experimental results indicate that the new algorithm is usually faster than earlier methods.

Key words. sparse matrix algorithms, Gaussian elimination, *LU* factorization, elimination tree, elimination dag

AMS(MOS) subject classifications. 05C20, 05C75, 65F05, 65F50.

1. Introduction.

The *elimination tree* [10], [14] is central to the study of symmetric factorization of sparse positive definite matrices. Liu [11] surveys the use of this tree structure in many aspects of sparse Cholesky factorization, including sparse storage schemes, matrix reordering, symbolic and numerical factorization algorithms, and modeling parallel elimination. In particular, the row and column structures of the Cholesky factor of a symmetric positive definite sparse matrix can be characterized in terms of its elimination tree.

The objective of this paper is to generalize the elimination tree to a structure that can be used to study sparse unsymmetric *LU* factorization. We seek a structure that characterizes the lower and upper triangular factors in the same way that elimination trees characterize Cholesky factors. Moreover, for the special case when the matrix is symmetric, the generalized structure should be the elimination tree. Our generalization consists of a pair of special directed acyclic graphs (dags), which we call the *elimination dags*, or *edags* for short.

The outline of the paper is as follows. In §2, we provide the background material to study sparse *LU* factorization. We formulate numerical factorization as a sequence of sparse triangular solves, thus motivating the study of the solution of sparse triangular systems. We introduce the relevant graph notions, and quote results from the literature that relate the structure of the solution vector to the structures of the triangular matrix and the right-hand side vector.

In §3, we define elimination dags. They are simply the smallest structures that preserve the set of paths in the graphs of the lower and upper triangular factor matrices. In graph-theoretic terms, the edags are *transitive reductions* [1] of the graphs of L and U. We show that if the matrix is symmetric, elimination dags are simply elimination trees. We prove some graph-theoretic results about the path structure of the graph of a matrix, its triangular factors, and its elimination dags.

[*] Received by the editors March 27, 1991; accepted for publication (in revised form) October 8, 1991.

[†] Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, California 94304 (gilbert@parc.xerox.com).

[‡] Department of Computer Science, York University, North York, Ontario, Canada M3J 1P3 (joseph@cs.yorku.ca). This research was supported in part by the Canadian Natural Sciences and Engineering Research Council under grant A5509.

Section 4 contains the main results, which characterize the structures of sparse LU factors in terms of elimination dags. We demonstrate that elimination dags are truly analogs of elimination trees, by showing that the symmetric structure theory of elimination trees can be obtained as special cases of the elimination dag results.

In §5, we apply elimination dags to sparse symbolic LU factorization. We briefly review two algorithms, FILL1 and FILL2, by Rose and Tarjan [12] that compute fill for sparse unsymmetric matrices. We then formulate a new symbolic factorization scheme based on the pair of elimination dags and compare its performance with that of FILL1 and FILL2. Our experimental results show that the new algorithm performs much better than either of these on practical sparse problems.

Section 6 contains our concluding remarks. We discuss extensions to these ideas, including the possibility of using elimination dags in numerical sparse LU factorization schemes with partial pivoting.

2. Background.

2.1. *LU* factorization in terms of triangular solutions. Suppose that A is an unsymmetric sparse $n \times n$ matrix that has a factorization $A = LU$ without pivoting. The *bordering method* allows us to compute the factorization as a sequence of triangular solves. Specifically, write

$$A = \begin{pmatrix} A' & h \\ g^T & s \end{pmatrix},$$

where A' is the $(n-1) \times (n-1)$ leading principal submatrix, g and h are $(n-1)$-vectors, and s is a scalar. If A' has been factored as $L'U'$, the LU factorization of A is

$$A = \begin{pmatrix} A' & h \\ g^T & s \end{pmatrix} = \begin{pmatrix} L' & \\ \bar{g}^T & \bar{s} \end{pmatrix} \begin{pmatrix} U' & \bar{h} \\ & 1 \end{pmatrix},$$

where $\bar{g} = U'^{-T} g$, $\bar{h} = L'^{-1} h$, and $\bar{s} = s - \bar{g}^T \bar{h}$.

Note that the vectors $\bar{g}$ and $\bar{h}$ can be obtained as the solutions of two lower triangular systems, each of size $n-1$:

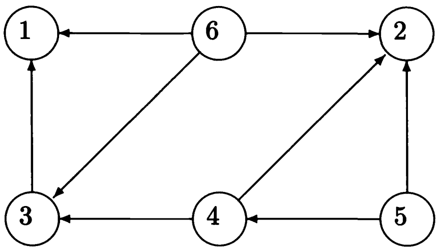
$$L'\bar{h} = h, \qquad U'^T \bar{g} = g.$$

The matrix A' can be factored into $L'U'$ in the same way recursively. Thus we can view the entire factorization as a sequence of n pairs of triangular solves of sizes from 0 to $n-1$. Moreover, the structure of the vector $\bar{h}$ depends on that of h and L', and the structure of $\bar{g}$ depends on that of g and U'. This motivates the following study relating the structure of the solution vector with that of the given lower triangular matrix and right-hand side.

2.2. Sparse triangular matrices and directed acyclic graphs. Let $A = (a_{ij})$ be a sparse $n \times n$ matrix with nonzero diagonal entries. Define $G(A)$ to be the directed graph of the matrix A as follows. The vertex set of $G(A)$ is $V = \{1, 2, \ldots, n\}$; there is an edge from i to j (for $i \neq j$) if and only if the entry a_{ij} is nonzero. We write the directed edge from i to j as $\langle i, j \rangle$. Note that the edges $\langle i, j \rangle$ and $\langle j, i \rangle$ are different; the former means that $a_{ij} \neq 0$ and the latter means that $a_{ji} \neq 0$.

Many sparse matrix codes [6], [9] use a data structure that lists the nonzeros of a matrix A in column major order. In graph terms, this is the "adjacency list" structure for $G(A^T)$. Representing A by $G(A^T)$ instead of by $G(A)$ seems to be necessary for

$$L = \begin{pmatrix} 1 & & & & & \\ & 2 & & & & \\ \bullet & & 3 & & & \\ & \bullet & \bullet & 4 & & \\ & \bullet & & \bullet & 5 & \\ \bullet & \bullet & \bullet & & & 6 \end{pmatrix}$$



$G(L)$

FIG. 1. *A lower triangular matrix L and its directed graph $G(L)$.*

some efficient algorithms [9]. This is the reason that this paper sometimes states results and algorithms in terms of the graph of the transpose of the matrix in question.

The directed graph of a triangular matrix has a special structure. Consider a lower triangular matrix $L = (\ell_{ij})$ and its directed graph $G(L)$. A nonzero off-diagonal entry ℓ_{ij} must have $i > j$, so any directed edge $\langle i, j \rangle$ of $G(L)$ must satisfy $i > j$. Since a directed edge always points from a vertex with a higher subscript to one with a lower subscript, the graph $G(L)$ must be *acyclic*: it cannot have any directed cycles. An acyclic directed graph is often called a *dag* for short [2]. Figure 1 is an example of a lower triangular matrix and its corresponding directed acyclic graph.

2.3. Structural characterization of triangular solution. Consider the solution of the lower triangular system $Lx = b$, where both L and the right-hand side vector b are sparse. Gilbert [7] provides a simple characterization of the sparse structure of the solution vector x in terms of that of L and b. We introduce the following notation: For any row or column vector $w = (w_1, \ldots, w_n)$ or $w = (w_1, \ldots, w_n)^T$, define its *vector structure* as the vertex subset

$$\text{Struct}(w) = \{i \in V \mid w_i \neq 0\}.$$

Note that in this definition, i is a vertex in the graph $G(L)$ and w_i is an entry in the vector w.

THEOREM 2.1 (see [7]). *The structure* Struct(x) *of the solution vector x to $Lx = b$ is given by the set of vertices reachable from vertices of the right-hand side structure* Struct(b) *by paths in the directed graph $G(L^T)$.*

Strictly speaking, this result gives only an upper bound on the structure of x, because coincidental cancellation might produce more zeros. This is the tightest upper bound on Struct(x) that is possible given only the nonzero structures of L and b. To avoid cluttering up theorem statements in the rest of the paper, we will not keep mentioning the possibility of coincidental cancellation. Thus when we make a statement of the form, "The structure of X is Y," it should be understood to mean all of the following: "If the nonzero values of A are chosen independently at random, then the structure of X is Y with probability one"; "If the nonzero values of A are algebraically independent, then the structure of X is Y"; and "Regardless of the nonzero values of A, the structure of X is a subset of Y." See Gilbert [7] for a detailed discussion of this point.

To illustrate Theorem 2.1, consider the solution of $Lx = b$ for the matrix L of Fig. 1. The graph $G(L^T)$ is the graph $G(L)$ with all its edges reversed. If b_1 is the only nonzero in the right-hand side vector b, then its structure vertex set is Struct$(b) = \{1\}$.

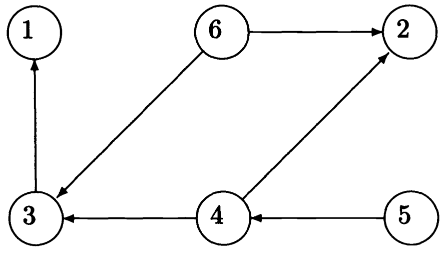


FIG. 2. *The transitive reduction of the graph in Fig.* 1.

The vertices that are reachable from vertex 1 through directed paths in $G(L^T)$ are those in the set $\{1,3,4,5,6\}$. On the other hand, if Struct$(b) = \{2,4\}$, then the solution structure Struct(x) is $\{2,4,5,6\}$.

Theorem 2.1 is actually true (including the remarks about coincidental cancellation) for any nonsingular matrix with nonzero diagonal, whether triangular or not. However, we will only use the triangular case.

3. Elimination dags for sparse *LU* factorization.

3.1. Transitive reduction of a directed graph. Theorem 2.1 relates the structure of the solution vector x to information about paths in the directed graph $G(L)$. One economical way to represent path information for a directed graph is by its *transitive reduction* [1]. The idea is to find another directed graph with fewer edges than the given graph $G(L)$, but with the same path structure. In such a graph, fewer edges would need to be traversed to generate Struct(x) for a given Struct(b).

A graph G° is a *transitive reduction* of a given directed graph G if G° satisfies the following two conditions:

(a) G° has a directed path from u to v if and only if G has a directed path from u to v.

(b) No graph with fewer edges than G° satisfies condition (a).

An arbitrary graph may have many different transitive reductions, but Aho, Garey, and Ullman [1] show that a dag has only one.

THEOREM 3.1 (see [1]). *If G is a directed acyclic graph, then the transitive reduction of G is unique, and it is a subgraph of G.*

Figure 2 shows the transitive reduction of the graph in Fig. 1. Since there is a directed path $\langle 6,3,1\rangle$, the directed edge $\langle 6,1\rangle$ is redundant. Similarly, $\langle 5,2\rangle$ is redundant because of the path $\langle 5,4,2\rangle$.

Since the transitive reduction preserves paths, we can restate Theorem 2.1 in terms of it.

COROLLARY 3.2. *The structure* Struct(x) *of the solution vector x to $Lx = b$ is given by the set of vertices reachable from vertices of the right-hand side structure* Struct(b) *by paths in the transitive reduction $G^\circ(L^T)$ of the directed graph $G(L^T)$.*

A notion related to transitive reduction is the *transitive closure* G^* of a directed graph G, which is the graph that has an edge $\langle u,v\rangle$ whenever G has a directed path from u to v. The transitive closure is the least compact representation of the path information in G; the transitive reduction G° can also be defined as the smallest graph with the same transitive closure as G. Gilbert [7] shows that the transitive closure captures the structure of the inverse of a matrix:

THEOREM 3.3 (see [7]). *Let A be a nonsingular matrix, not necessarily triangular, with nonzero diagonal elements. Ignoring coincidental cancellation, $G(A^{-1}) = G^*(A)$.*

$$A = \begin{pmatrix} 1 & & \bullet & & & \\ & 2 & & & & \\ \bullet & & 3 & & & \bullet \\ & \bullet & \bullet & 4 & \bullet & \\ & \bullet & & \bullet & 5 & \\ \bullet & \bullet & & & & 6 \end{pmatrix} = \begin{pmatrix} 1 & & & & & \\ & 2 & & & & \\ \bullet & & 3 & & & \\ & \bullet & \bullet & 4 & & \\ & \bullet & & \bullet & 5 & \\ \bullet & \bullet & \circ & & & 6 \end{pmatrix} \begin{pmatrix} 1 & & \bullet & & & \\ & 2 & & & & \\ & & 3 & & & \bullet \\ & & & 4 & \bullet & \circ \\ & & & & 5 & \circ \\ & & & & & 6 \end{pmatrix}$$

FIG. 3. *LU factorization of a matrix example.*

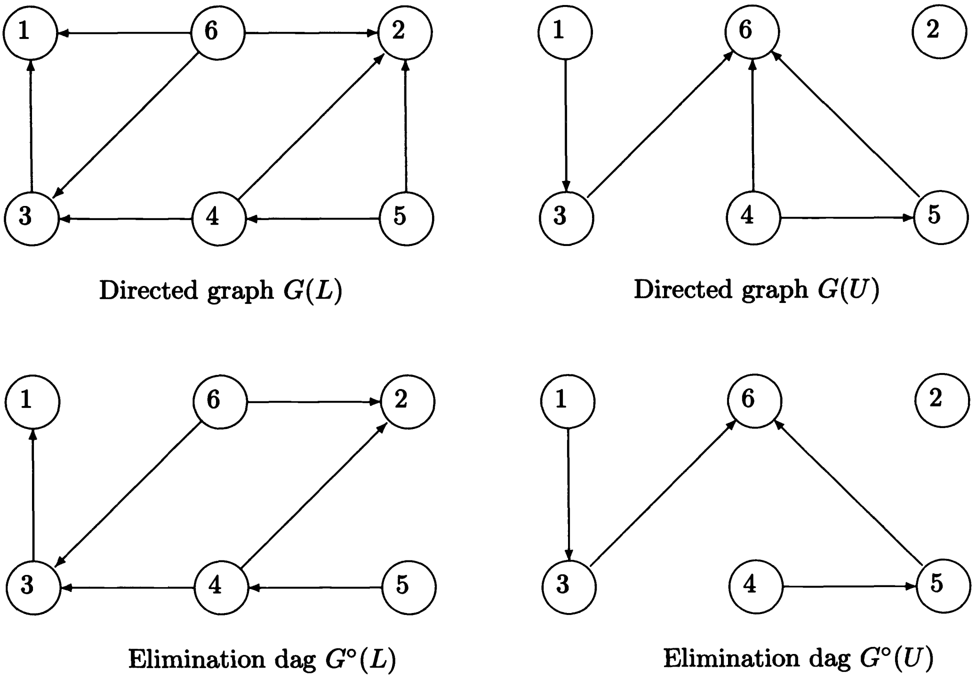


FIG. 4. *Elimination dags of the matrix in Fig. 3.*

3.2. Definition of elimination dags. As before, let A be an unsymmetric $n \times n$ matrix that can be factored as $A = LU$ without pivoting. The lower triangular matrix L has a directed acyclic graph $G(L)$, which has a unique transitive reduction $G^\circ(L)$ by Theorem 3.1.

Similarly, the upper triangular matrix U has a directed acyclic graph $G(U)$, which in turn has a unique transitive reduction $G^\circ(U)$. We call the two reduced directed acyclic graphs $G^\circ(L)$ and $G^\circ(U)$ the *lower* and *upper elimination directed acyclic graphs*, respectively. We also refer to them collectively as the *elimination dags* of the matrix A.

As an illustration, consider the unsymmetric matrix in Fig. 3. An LU factorization of this matrix creates three fills, one in L and two in U. The fills are represented in the figure by "o." Figure 4 displays the two elimination dags corresponding to this matrix example.

3.3. The symmetric case: Elimination trees. The situation is simpler if the matrix A is symmetric. Still assuming no pivoting, Gaussian elimination yields a symmetric factorization $A = LL^T$. The *elimination tree* [10], [14] has been used extensively to study symmetric Gaussian elimination.

Let L be the Cholesky factor of a symmetric positive definite matrix A. Formally, the elimination tree $T(A)$ of A has n vertices $\{1, \dots, n\}$, and $\langle j, k\rangle$ is an edge if and only if

$$k = \min\{r > j \mid \ell_{rj} \neq 0\}.$$

This structure is a tree rooted at vertex n if A is irreducible, or a forest with one tree for each irreducible block if A is reducible. Liu's survey paper [11] contains a comprehensive exposition of this important structure in the context of sparse factorization. We quote one of its properties in our terminology.

THEOREM 3.4 (see [11]). *The elimination tree $T(A)$ is the transitive reduction of the directed graph $G(L^T)$.*

Since the transitive reduction of a dag is unique, it follows that for a symmetric matrix the elimination dags $G^\circ(L^T)$ and $G^\circ(U)$ are both equal to the elimination tree $T(A)$. In §4, we shall show that the structural characterization of Cholesky factors by elimination trees can be generalized to unsymmetric factors using elimination dags.

3.4. Elimination dags and path structure. Elimination dags capture the path structure of the LU factors of a sparse matrix. In this section we explore some of the connections between edags, the path structures of L and U, and the path structure of the original matrix A. We also answer the question, "Which pairs of directed graphs can be the elimination dags of some matrix?" Recall that we consider only matrices A with nonzero diagonal elements, and recall the discussion of coincidental cancellation in §2.3.

We begin by defining notation for two directed graphs that correspond to addition and multiplication of matrices.

DEFINITION. If B and C are two $n \times n$ matrices with nonzero diagonal elements, then $G(B) + G(C)$ is the *union* of the graphs of B and C, defined as the graph whose vertex set is $\{1, \dots, n\}$ and whose edge set is the union of those of $G(B)$ and $G(C)$.

DEFINITION. If B and C are two $n \times n$ matrices with nonzero diagonal elements, then $G(B) \cdot G(C)$ is the *product* of the graphs of B and C, defined as the graph whose vertex set is $\{1, \dots, n\}$, with an edge $\langle i, j\rangle$ exactly in case $\langle i, j\rangle$ is an edge of $G(B)$, or $\langle i, j\rangle$ is an edge of $G(C)$, or there is some k such that $\langle i, k\rangle$ is an edge of $G(B)$ and $\langle k, j\rangle$ is an edge of $G(C)$.

We will write $G \subseteq H$ to mean that G is a subgraph of H.

LEMMA 3.5. *If B and C are matrices with nonzero diagonals, then*

$$G(B + C) \subseteq G(B) + G(C),$$

with equality unless there is coincidental cancellation in $B + C$.

Proof. The proof follows immediately. □

LEMMA 3.6. *If B and C are matrices with nonzero diagonals, then*

$$G(BC) \subseteq G(B) \cdot G(C),$$

with equality unless there is coincidental cancellation in BC.

Proof. If $A = BC$, then $a_{ij} \neq 0$ only if there exists k such that $b_{ik} \neq 0$ and $c_{kj} \neq 0$. Conversely, if such k exists then a_{ij} is a sum of nonzeros and hence can be zero only if there is cancellation. □

If $A = LU$ is an LU factorization, Lemma 3.6 implies that $G(A)$ is a subgraph of $G(L) \cdot G(U)$, that is, that an edge in A corresponds to an edge in L followed by an edge in U. However, $G(A)$ is not usually equal to $G(L) \cdot G(U)$; any fill in L and U cancels out when they are multiplied back together.

Our first result relates paths in A to paths in L and U. It says that a path in A always corresponds to a path in U followed by a path in L. We allow paths of length zero.

THEOREM 3.7. *If $A = LU$ and there is a path in the directed graph $G(A)$ from vertex i to vertex j, then there is some $1 \leq k \leq n$ such that the directed graph $G(U)$ has a path from vertex i to vertex k, and the directed graph $G(L)$ has a path from vertex k to vertex j.*

Proof. This can be proved by induction using the "path lemma" of Rose and Tarjan [12], which characterizes $G(L+U)$ in terms of $G(A)$. Another proof, however, is to interpret the equation $A^{-1} = U^{-1}L^{-1}$ in terms of paths and transitive closures. Theorem 3.3 implies that

$$\begin{aligned} G^*(A) &= G(A^{-1}) \\ &= G(U^{-1}L^{-1}) \\ &\subseteq G(U^{-1}) \cdot G(L^{-1}) \\ &= G^*(U) \cdot G^*(L), \end{aligned}$$

where the third step is by Lemma 3.6. In English, this equation says that every path in A corresponds to a path in U followed by a path in L. □

Since the elimination dags of A preserve the path information in $G(L)$ and $G(U)$, the same conclusion holds for them.

COROLLARY 3.8. *If $A = LU$ and there is a path in the directed graph $G(A)$ from vertex i to vertex j, then there is some $1 \leq k \leq n$ such that the elimination dag $G^\circ(U)$ has a path from vertex i to vertex k, and the elimination dag $G^\circ(L)$ has a path from vertex k to vertex j. In other words,*

$$G^*(A) \subseteq G^{\circ *}(U) \cdot G^{\circ *}(L).$$

This says that, while the elimination dags have very simple (i.e., acyclic) path structures themselves, they preserve all the path structure of A. Any path in $G(A)$ corresponds to an "ascending" path in $G^\circ(U)$ followed by a "descending" path in $G^\circ(L)$.

We turn now to the question of characterizing the graphs that can be elimination dags of some matrix. Again we consider $G(L)$ and $G(U)$ first. Any directed graph G with vertices $\{1, \ldots, n\}$ whose edges $\langle i, j \rangle$ all satisfy $i > j$ is $G(L)$ for some A. We merely choose A to be the matrix whose entries are $a_{ij} = 1$, whenever $i = j$ or $\langle i, j \rangle$ is an edge of G, and zeros elsewhere. Then $A = LU$ with $L = A$ and $U = I$, and $G(L) = G(A) = G$. Similarly any graph with edges directed from lower to higher numbered vertices is the graph of U for some A.

However, not every *pair* of directed acyclic graphs corresponds to a factorization $A = LU$ without cancellation. Rose and Tarjan [12] show that (in the absence of coincidental cancellation) $G(L+U)$ is always a so-called *perfect elimination digraph* in perfect elimination order, which means that whenever $k < \min(i, j)$ and $\langle i, k \rangle$ and $\langle k, j \rangle$ are edges, then $\langle i, j \rangle$ is also an edge. This condition is both necessary and sufficient. We can restate it in terms of our definitions as follows.

Theorem 3.9 (see [12]). *Suppose there is no cancellation in the factorization* $A = LU$. *Then*

$$G(L) \cdot G(U) = G(L) + G(U). \tag{1}$$

Furthermore, a specified lower and upper triangular structure $G(L)$ *and* $G(U)$ *correspond to the LU factorization without cancellation of some matrix if and only if they satisfy* (1).

The situation for structures of elimination dags is similar: Any transitively reduced dag can be the structure of the elimination dag $G^\circ(L)$ or $G^\circ(U)$, but a specified pair of reduced dags might not be $G^\circ(L)$ and $G^\circ(U)$ for any single matrix $A = LU$. We can characterize the possible pairs in a way similar to Theorem 3.9.

Theorem 3.10. *Suppose there is no cancellation in the factorization* $A = LU$. *Then*

$$G^{\circ *}(L) \cdot G^{\circ *}(U) = G^{\circ *}(L) + G^{\circ *}(U). \tag{2}$$

Furthermore, a reduced structure is the pair of edags of some matrix if and only if it satisfies (2).

Remark. An equivalent way to state the theorem is: Let two transitively reduced acyclic directed graphs G_ℓ and G_u be given, with edges oriented from higher- to lower-numbered vertices in G_ℓ and vice versa in G_u. Then there exists a matrix A that factors without cancellation as LU with $G^\circ(L) = G_\ell$ and $G^\circ(U) = G_u$ if and only if $G_\ell^* \cdot G_u^* = G_\ell^* + G_u^*$.

Proof. First, let $A = LU$ be given. The containment $G^{\circ *}(L) + G^{\circ *}(U) \subseteq G^{\circ *}(L) \cdot G^{\circ *}(U)$ follows from the definition of the product graph. We prove the opposite containment. Note that $G^{\circ *}(L) = G^*(L)$ and $G^{\circ *}(U) = G^*(U)$.

Let $\langle i, j \rangle$ be an edge of $G^{\circ *}(L) \cdot G^{\circ *}(U)$. Then there exists k such that there are paths $\langle i = i_t, i_{t-1}, \ldots, i_0 = k \rangle$ in $G(L)$ and $\langle k = j_0, j_1, \ldots, j_s = j \rangle$ in $G(U)$. We induct on $s + t$, the total length of these paths. If either path has length zero, then the concatenation of the paths is in $G(L)$ or $G(U)$ and $\langle i, j \rangle$ is an edge of $G^*(L) + G^*(U) = G^{\circ *}(L) + G^{\circ *}(U)$. If both paths have length at least one and $i_1 = j_1$, then we can shorten the paths by deleting $i_0 = j_0$. Otherwise, the edges $\langle i_1, k \rangle$ in $G(L)$ and $\langle k, j_1 \rangle$ in $G(U)$ make $\langle i_1, j_1 \rangle$ an edge of $G(L) \cdot G(U)$. Theorem 3.9 then implies that $\langle i_1, j_1 \rangle$ is an edge of $G(L) + G(U)$; that is, it is either an edge of $G(L)$ or an edge of $G(U)$. If $i_1 > j_1$ then $\langle i_1, j_1 \rangle$ is an edge of $G(L)$. Then the paths $\langle i = i_t, i_{t-1}, \ldots, i_1, j_1 \rangle$ in $G(L)$ and $\langle j_1, j_2 \ldots, j_s = i \rangle$ in $G(U)$ have smaller total length than the original paths, so the inductive hypothesis applies. Similarly, if $i_1 < j_1$ then we shorten the path in $G(L)$.

Second, let G_ℓ and G_u be given. The "only if" part of the theorem is what we have just proved. To prove the "if" part, we assume that $G_\ell^* \cdot G_u^* = G_\ell^* + G_u^*$ and proceed to show the existence of a suitable A. Now G_ℓ^* and G_u^* satisfy (1), so Theorem 3.9 says that there exists $A = LU$ (without cancellation) such that $G(L) = G_\ell^*$ and $G(U) = G_u^*$. But since G_ℓ and G_u are transitively reduced and the transitive reduction of a dag is unique, this implies $G^\circ(L) = G_\ell$ and $G^\circ(U) = G_u$ as well. □

All these results reduce to fairly simple facts about elimination trees in the symmetric case. In the symmetric version of Theorem 3.7, an irreducible A gives a complete graph $G^*(A)$ and a complete lower triangular structure $G^*(L)$. Theorem 3.9 says in the symmetric case that $G(L) \cdot G(L^T) = G(L) + G(L^T)$, which restates that $G(L)$ is a perfect elimination or *chordal* graph [13]. The symmetric case of Theorem 3.10 is an unusual characterization of a forest (that is, of a graph whose connected components are trees): $G^{\circ *}(L) \cdot G^{\circ *}(L^T) = G^{\circ *}(L) + G^{\circ *}(L^T)$ says that a "descending" path in $T(A)$ followed by an "ascending" path in $T(A)$ can be replaced by a pure descending or ascending path. This is indeed true of forests, and forests are the only transitively reduced directed acyclic graphs of which it is true.

The latter observation is a roundabout proof that the elimination tree of a symmetric matrix really is a tree (or, rather, a forest). One important fact about an elimination tree of an $n \times n$ matrix is that it encodes all the path information about $G(L)$ in a structure of linear size—with n vertices and at most $n-1$ edges. Thus $G^\circ(L)$ is likely to be much smaller than L.

An elimination dag, on the other hand, may have $\Theta(n^2)$ edges in the worst case. Matrix EG-3 in Fig. 10 is an example where both $G(U)$ and $G^\circ(U)$ have about $n^2/4$ edges. Nevertheless, in practice we observe that the elimination dags of a matrix are almost always much smaller than its triangular factors. This fact explains the results of §5.3, in which an algorithm based on edags is much faster than two earlier algorithms.

4. Structural characterization of *LU* factors.

In this section we use the elimination dags $G^\circ(L)$ and $G^\circ(U)$ to characterize the row and column structures of the triangular factors. Throughout the section, $A = LU$ is the factorization without pivoting of a square matrix A with nonzero diagonal. Recall the discussion in §2.3 of our assumptions about cancellation.

4.1. Row structure of *L*.

We first focus our attention on the row structure of the lower triangular factor L. The definitions immediately imply the following proposition.

PROPOSITION 4.1. *If $\ell_{ij} \neq 0$, then there exists a path from vertex i to vertex j in the elimination dag $G^\circ(L)$.*

This condition is necessary but clearly not sufficient for an entry in L to be nonzero. For example, the entry ℓ_{41} is zero in Fig. 3, but there is a path from vertex 4 to 1 via 3. We now provide a necessary and sufficient condition for ℓ_{ij} to be nonzero in terms of paths in the upper elimination dag $G^\circ(U)$.

THEOREM 4.2. *Let $i > j$. Then $\ell_{ij} \neq 0$ if and only if there is some $k \leq j$ such that $a_{ik} \neq 0$ and there is a directed path in the elimination dag $G^\circ(U)$ from vertex k to vertex j.*

Proof. Let A_i be the $i \times i$ leading principal submatrix of A, with factors $A_i = L_i U_i$. Then the ith row of L (without the diagonal entry) is given by

$$(\ell_{i1}, \ldots, \ell_{i\,i-1})^T = \begin{pmatrix} \ell_{i1} \\ \vdots \\ \ell_{i\,i-1} \end{pmatrix} = U_{i-1}^{-T} \begin{pmatrix} a_{i1} \\ \vdots \\ a_{i\,i-1} \end{pmatrix} = U_{i-1}^{-T}(a_{i1}, \ldots, a_{i\,i-1})^T.$$

Therefore, by Theorem 2.1, the vector structure of $(\ell_{i1}, \ldots, \ell_{i,i-1})$ is given by the set of vertices reachable from the vertices in the vector structure of $(a_{i1}, \ldots, a_{i,i-1})$ in the directed graph $G(U_{i-1})$. Thus $\ell_{ij} \neq 0$ if and only if there exists a nonzero a_{ik} such that vertex j is reachable from vertex k in the graph $G(U_{i-1})$, in other words, if and only if there exists k such that $a_{ik} \neq 0$ and $G(U_{i-1})$ contains a directed path from k to j. Such a k must be less than or equal to j because all edges of $G(U)$ are directed from lower- to higher-numbered vertices.

Now $G(U_{i-1})$ is a subgraph of $G(U)$, so a path in $G(U_{i-1})$ is a path in $G(U)$ as well. Conversely (again because edges are directed from lower- to higher-numbered

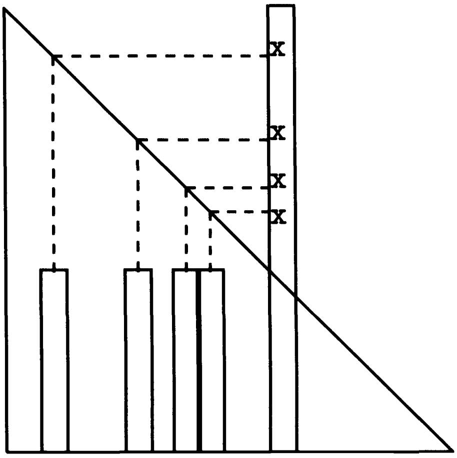


FIG. 5. *Union of column structures.*

vertices) any directed path ending at vertex j in $G(U)$ must lie entirely within $G(U_{i-1})$. Thus $\ell_{ij} \neq 0$ if and only if there exists k such that $a_{ik} \neq 0$ and there is a directed path from k to j in $G(U)$. The result now follows from the fact that the elimination dag $G^{\circ}(U)$ preserves the set of paths in $G(U)$. □

Theorem 4.2 characterizes the structure of L by rows: the structure of the ith row of L is given by the subset of $\{1,\ldots,i\}$ reachable in the upper elimination dag $G^{\circ}(U)$ from the vertices of the structure of the ith row in the lower triangular part of A. The following section characterizes the structure of L by columns.

4.2. Column structure of L. Using the notation $\mathrm{Struct}(L_{*j}) = \{i \mid \ell_{ij} \neq 0\}$ for the nonzero structure of a column of L, we restate a result of Rose and Tarjan [12] that describes the structure of L by columns in terms of A and U.

THEOREM 4.3 (see [12]).

$$\begin{aligned}&\mathrm{Struct}(L_{*j}) = \\ &\quad \mathrm{Struct}(A_{*j}) \cup \bigcup\{\mathrm{Struct}(L_{*k}) \mid k < j, u_{kj} \neq 0\} - \{1,\ldots,j-1\}.\end{aligned}$$

COROLLARY 4.4. *Let $k < j$ and $u_{kj} \neq 0$. Then*

$$\mathrm{Struct}(L_{*k}) - \{1,\ldots,j-1\} \subseteq \mathrm{Struct}(L_{*j}).$$

In matrix terms, we can determine the column structure of L_{*j} by taking the union of the column structure of A_{*j} with a number of columns of L before j. These columns are given exactly by the structure of column j of U, the upper triangular factor. Figure 5 expresses the result of Theorem 4.3 diagrammatically.

Note that all columns L_{*k} such that $u_{kj} \neq 0$ are used to determine the structure of L_{*j} in Theorem 4.3. The following theorem shows that we need only consider a subset of those columns, and this subset is determined by the structure of the upper elimination dag $G^{\circ}(U)$.

THEOREM 4.5.

$$\text{Struct}(L_{*j}) =$$
$$\text{Struct}(A_{*j}) \cup \bigcup\{\text{Struct}(L_{*k}) \mid \langle k, j\rangle \textit{ is an edge of } G^\circ(U)\} - \{1, \ldots, j-1\}.$$

Proof. The right-hand side in Theorem 4.5 is a subset of the right-hand side of Theorem 4.3 by the definition of $G^\circ(U)$.

To prove the converse, consider any nonzero entry u_{kj} where $\langle k, j\rangle$ is not an edge of $G^\circ(U)$. Then $G^\circ(U)$ must contain a path $\langle k = i_0, i_1, \ldots, i_t, i_{t+1} = j\rangle$, with $u_{i_s i_{s+1}} \neq 0$ for $0 \leq s \leq t$. Corollary 4.4 says that

$$\begin{aligned}
\text{Struct}(L_{*k}) - \{1, \ldots, i_1 - 1\} &\subseteq \text{Struct}(L_{*i_1}), \\
\text{Struct}(L_{*i_1}) - \{1, \ldots, i_2 - 1\} &\subseteq \text{Struct}(L_{*i_2}), \\
&\vdots \\
\text{Struct}(L_{*i_{t-1}}) - \{1, \ldots, i_t - 1\} &\subseteq \text{Struct}(L_{*i_t}).
\end{aligned}$$

Combining, we have

$$\text{Struct}(L_{*k}) - \{1, \ldots, j-1\} \subseteq \text{Struct}(L_{*i_t}) - \{1, \ldots, j-1\},$$

where $\langle i_t, j\rangle$ is an edge of $G^\circ(U)$. We have thus proved that the right-hand side in Theorem 4.3 is a subset of that in Theorem 4.5. □

4.3. Row and column structures of U. By the same argument as in Theorems 4.2 and 4.5, we can relate the nonzero structure of the upper triangular factor U to the lower elimination dag $G^\circ(L)$. Theorem 4.6 characterizes the column structure, and Theorem 4.7 the row structure, of U.

THEOREM 4.6. *Let $i < j$. Then $u_{ij} \neq 0$ if and only if there is some $k \leq i$ such that $a_{kj} \neq 0$ and there is a directed path in the elimination dag $G^\circ(L)$ from vertex i to vertex k.*

THEOREM 4.7.

$$\text{Struct}(U_{i*}) =$$
$$\text{Struct}(A_{i*}) \cup \bigcup\{\text{Struct}(U_{k*}) \mid \langle i, k\rangle \textit{ is an edge of } G^\circ(L)\} - \{1, \ldots, i-1\}.$$

We now use the matrix example in Fig. 3 to illustrate the results of Theorems 4.2 and 4.6 on the row structure of L and the column structure of U. Consider row and column 6 of the matrix in Fig. 3. The structure of row A_{6*} is $\{1, 2, 6\}$. The set of vertices reachable from this set in the upper elimination dag $G^\circ(U)$ in Fig. 4 is $\{1, 2, 3, 6\}$, which is precisely the structure of L_{6*}. On the other hand, the structure of column A_{*6} is $\{3, 6\}$. These two vertices in the lower elimination dag $G^\circ(L)$ are reachable exactly from vertices in the set $\{3, 4, 5, 6\}$. As Theorem 4.6 predicts, this is the structure of U_{*6}.

Using the same matrix, we illustrate the results of Theorems 4.5 and 4.7 on the column structure of L and the row structure of U. Consider row and column 5 of the matrix in Fig. 3. Since $\langle 4, 5\rangle$ is the only edge into vertex 5 in the upper elimination dag $G^\circ(U)$, Theorem 4.5 says that $\text{Struct}(L_{*5})$ can be obtained from $\text{Struct}(A_{*5})$ and $\text{Struct}(L_{*4})$. On the other hand, of the two edges $\langle 5, 2\rangle$ and $\langle 5, 4\rangle$ in the graph $G(L)$, only $\langle 5, 4\rangle$ is in the lower elimination dag $G^\circ(L)$. Therefore, by Theorem 4.7, it is sufficient to consider $\text{Struct}(A_{5*})$ and $\text{Struct}(U_{4*})$ to determine $\text{Struct}(U_{5*})$.

4.4. Edags as unsymmetric analogs of elimination trees. In §3.3 we saw that when the matrix A is symmetric, the elimination dag $G^\circ(L^T) = G^\circ(U)$ is identical to the elimination tree $T(A)$. Here we show that the results from the literature relating elimination trees to the structure of sparse Cholesky factors are special cases of the results of §§4.1–4.3. Thus elimination dags are truly an unsymmetric analog of elimination trees.

In this subsection, we take A to be symmetric and positive definite, and L to be its Cholesky factor. The following result characterizes the column structure of L using the elimination tree, and forms the basis of an efficient symbolic Cholesky factorization scheme.

THEOREM 4.8 (see [11]).

$$\text{Struct}(L_{*j}) =$$
$$\text{Struct}(A_{*j}) \cup \bigcup \{\text{Struct}(L_{*k}) \mid k \textit{ is a child of } j \textit{ in } T(A)\} - \{1, \ldots, j-1\}.$$

Since $G^\circ(L^T) = T(A)$ in the symmetric case, this is a special case of Theorem 4.5 or Theorem 4.7.

The row structure of the Cholesky factor L can also be characterized in terms of the elimination tree. Schreiber [14] shows that the row structure of L_{i*} is a pruned subtree rooted at vertex i of the elimination tree. Liu [10] gives a complete characterization, which we quote here.

THEOREM 4.9 (see [10]). *Let $i > j$. Then $\ell_{ij} \neq 0$ if and only if vertex j is an ancestor of some vertex k in the elimination tree such that $a_{ik} \neq 0$.*

We can restate this in terms of paths in the elimination tree as follows.

COROLLARY 4.10. *Let $i > j$. Then $\ell_{ij} \neq 0$ if and only if there is some $k \leq j$ such that $a_{ik} \neq 0$ and there is a directed path in the elimination tree $T(A)$ from vertex k to vertex j.*

This is a special case of Theorem 4.2 or Theorem 4.6.

5. An application: Fill computation.

In this section we apply elimination dags to computing the symbolic factorization of an unsymmetric matrix. For this problem, we are given the nonzero structure of an unsymmetric matrix A and asked to compute the nonzero structures of its LU factors, under the assumption that A has an LU factorization without pivoting. Equivalently, we want to compute the result of playing the vertex elimination game [12] on a specified directed graph. We first review two algorithms from the literature briefly. Then we present a new algorithm, analyze it, and give experimental results.

In this section we use the notation $\eta(X)$ for the number of nonzeros in the matrix or vector X.

5.1. Algorithms FILL1 and FILL2. Rose and Tarjan [12] present two algorithms, which they call FILL1 and FILL2, to determine fill. Both algorithms compute the structures of L and U row by row.

Algorithm FILL1 determines the row structures of L_{i*} and U_{i*} by using the structure of A_{i*} together with the first $i-1$ computed row structures of U. It basically performs a symbolic simulation of the numerical row elimination scheme. The amount of work required is proportional to that required for numerical factorization. This is also proportional to the time required to multiply L and U back together again, which we write informally as flops(LU).

On the other hand, Algorithm FILL2 computes the row structures by using the "path lemma" for fills [12, Thm. 1], which says that $\langle i, j \rangle$ is an edge of $G(L+U)$ if

```
Algorithm EDAGS
  for row $i := 1$ to $n$ do
    1: Compute Struct($L_{i*}$) by traversing $G^\circ(U_{i-1})$
         from $A_{i*}$ (using Theorem 4.2);
    2: Transitively reduce Struct($L_{i*}$) using $G^\circ(L_{i-1})$,
         thus extending the edag to $G^\circ(L_i)$;
    3: Compute Struct($U_{i*}$) as a union of earlier rows of $U$
         (using Theorem 4.7);
    4: Transitively reduce Struct($U_{*i}$) using $G^\circ(U_{i-1})$,
         thus extending the edag to $G^\circ(U_i)$ ;
  end for
```

FIG. 6. *Symbolic LU factorization using elimination dags.*

and only if the graph $G(A)$ contains a directed path from i to j through vertices with numbers smaller than both i and j. The structures of L_{i*} and U_{i*} are determined using only the structures of the first i rows of the original matrix A. Rose and Tarjan show that the time complexity of FILL2 is bounded by $O(n\eta(A))$.

The reader is referred to the paper by Rose and Tarjan [12] for detailed descriptions of these algorithms. Rose and Tarjan do not provide code; our implementations, which we used to obtain the experimental results reported in this paper, are based on their descriptions.

5.2. A symbolic *LU* algorithm using elimination dags. The structural characterization results of Theorems 4.2, 4.5, 4.6, and 4.7 for sparse LU factors can be used to formulate a symbolic factorization algorithm. The algorithm in Fig. 6 computes the structures of L and U row by row. We use the notation L_i and U_i to denote the $i \times i$ leading principal submatrix of L and U. Their graphs and transitive reductions are defined in the same way as for L and U.

The algorithm uses Theorem 4.2 to determine the structure of L_{i*}, and Theorem 4.7 to determine the structure of U_{i*}. At the end of the ith iteration, the structures of the first i rows of L and U are known together with the structures of the elimination dags $G^\circ(L_i)$ and $G^\circ(U_i)$. The data structure represents L and A by columns, and U by rows; that is, it stores the adjacency lists of the various graphs $G(L^T)$, $G(U)$, $G^\circ(L^T)$, and $G^\circ(U)$. (The adjacency lists are not actually linked lists, but arrays of row numbers as in the standard column-oriented sparse matrix data structure [6].)

We now analyze the time complexity of this algorithm. To get simple expressions for the complexity, we will use the fact [9] that if there is no cancellation in solving $Lx = b$, then the number of nonzero arithmetic operations is the same as the number of nonzero arithmetic operations needed to multiply L by x, and is on the order of the time needed to traverse the part of the graph $G(L^T)$ reachable from b.

In step 1, at row i, this fact says that the traversal of $G^\circ(U_{i-1})$ takes time on the order of the number of operations to multiply a (hypothetical) matrix whose structure is that of $G^\circ(U_{i-1}^T)$ by the vector L_{i*}^T. Added up over all rows, this is bounded by the number of operations to multiply a matrix whose structure is that of $G^\circ(U^T)$ by L^T; we will write this number informally as flops($U^{\circ T}L^T$) or flops(LU°). Similarly, the time for step 3 is bounded by flops($L^\circ U$).

Step 2 can be implemented in at least two ways. One method is to search in

$G^\circ(L_{i-1})$ from the vertices corresponding to nonzeros in L_{i*}, reducing L_{i*} by discarding element ℓ_{ij} if vertex j is reached during the search from some ℓ_{ik} with $j < k$. The search at row i can avoid looking at any edge twice by searching in decreasing order of k and marking vertices that have already been reached. The time for this reduction search is on the order of the operations to solve for y in $L_{i-1}^{\circ T} y = L_{i*}^T$. If we add vertex i to get the graph $G^\circ(L_i)$, we see that the reduction search is the same as the search that would be done in solving $L_i^{\circ T} y = e_i$ (where e_i is the ith unit vector), except that from vertex i the reduction search examines all the nonzeros of L_{i*}^T instead of just those of $L_{i*}^{\circ T}$. Thus for row i, the search takes time on the order of operations to multiply $L_i^{\circ T}$ by column i of its inverse, plus $\eta(L_{i*}^T) - \eta(L_{i*}^{\circ T})$. Adding this up over all i, and noting that since $G(L)$ and $G^\circ(L)$ have the same transitive closures their inverses have the same structure, we get flops$(L^{-1}L^\circ) + \eta(L) - \eta(L^\circ)$. The last two terms are dominated by the first, so the total time for step 2 is on the order of flops$(L^{-1}L^\circ)$. Similarly, the total time for step 4 is on the order of flops$(U^\circ U^{-1})$ by this method.

The second method for implementing step 2 is to search "backwards" (that is, in $G^\circ(L_{i-1}^T)$), discarding nonzero ℓ_{ij} if vertex i can be reached from some ℓ_{ik} with $k < j$. Again the search can avoid looking at an edge twice by marking vertices that have been reached. The search at row i by this method is on the order of operations to solve $L_{i-1}^\circ y = L_{i*}^T$. It seems harder to get a simple expression for the total cost over all rows. To get an upper bound, we note that the search for row i at worst examines all of $G^\circ(L_{i-1})$, and thus the total time is no more than the order of $\sum_i \eta(L_i^\circ)$. The total time for step 4 is bounded above by a similar sum for U.

One way to express this upper bound is as follows. Let F be a full n by n "skew upper triangular" matrix, that is $f_{ij} \neq 0$ if $i+j \leq n+1$. Then $\sum_i \eta(L_i^\circ)$ is on the order of flops(FL°), and $\sum_i \eta(U_i^\circ)$ is on the order of flops$(U^\circ F)$. Thus the total time for the algorithm (using the second method for steps 2 and 4) is bounded by the order of flops$(FL^\circ + U^\circ F + LU^\circ + L^\circ U)$, which in turn is bounded above by flops$(F(E+E^T))$, where $E = L^\circ + U^\circ$ is a matrix with the structure of the two edags.

This upper bound actually applies to both methods for steps 2 and 4. In our implementation, we used the second method, since it uses $G^\circ(L^T)$ and thus fits more naturally with the column-oriented data structure for L.

Theoretically, none of the three algorithms FILL1, FILL2, or EDAGS dominates any of the others; any of them could be the fastest on some matrices. The next section describes three matrices, each of which takes $O(n^3)$ time by one algorithm and $O(n^2)$ time by the other two.

5.3. Experimental results. We compared the performance of the symbolic LU algorithm using elimination dags with the FILL1 and FILL2 algorithms of Rose and Tarjan. The programs were written in Fortran, using many of the routines from SPARSPAK [6]. We used sparse MATLAB [8] for the reorderings and to plot results. Figures 7, 8, and 9 show silhouette plots of the structure of a small sample matrix, its factors, and its elimination dags.

All but three of our test problems are from the Harwell–Boeing Sparse Matrix Collection [3]; Table 1 describes these problems. These problems all have unsymmetric nonzero structures, except for DWT0918, which is a symmetric problem that we included to demonstrate that EDAGS works well in that case too. We also included examples contrived to make each of the three algorithms perform poorly. Figure 10 shows 8×8 versions of these artificial problems; in the experiments we used 500×500 versions.

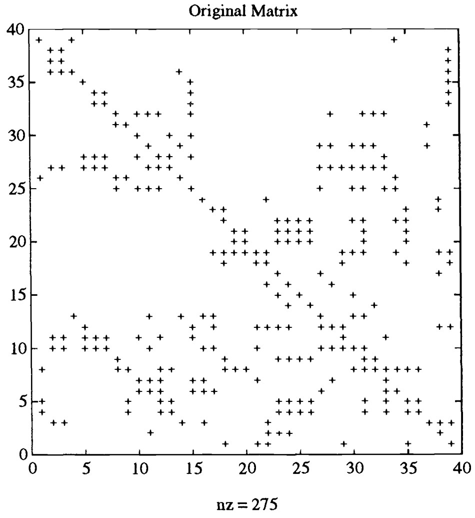


FIG. 7. *The nonzero structure of matrix A.*

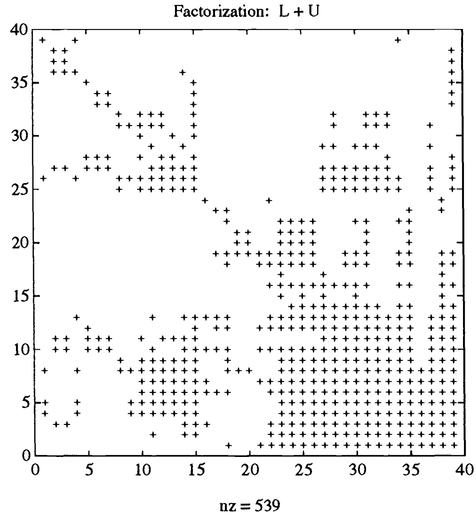


FIG. 8. *The triangular factors of $A = LU$.*

Since the algorithms are all nonnumerical and we are addressing only the case when no pivoting for numerical stability is necessary, the question of numerical error does not arise in the comparisons between the algorithms. We did, however, compute a numerical LU factorization for each matrix (after modifying the diagonal to make it dominant) in order to check that the combinatorial algorithms were producing the correct result.

Table 2 presents timing statistics for the three fill computation algorithms on the set of test problems. The experiments were performed on a Sun Sparcstation IPC workstation; times are reported in seconds as returned by calls to the system timer. For the Harwell–Boeing problems, FILL2 is consistently better than FILL1, whereas

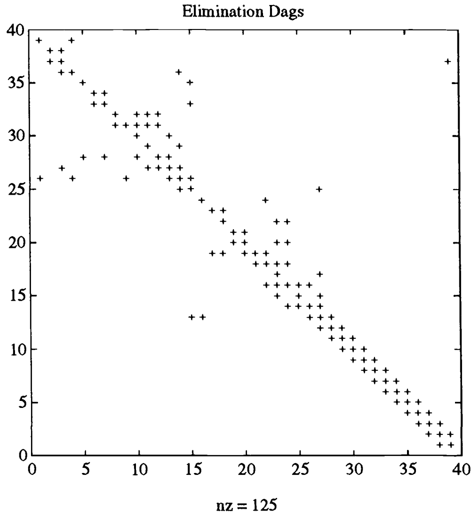


FIG. 9. *The elimination dags of A, $G^\circ(L) + G^\circ(U)$.*

TABLE 1
List of test problems.

Problem	n	Nonzeros	Description
SHL400	663	1712	Permuted triangular LP basis
FS7603	760	5976	Unsymmetric facsimile convergence matrix
MCFE	765	24382	Astrophysics problem
BP1600	822	4841	Unsymmetric LP basis
YOUNG3C	841	3988	Matrix from Young
DWT918	918	7384	Beam with cutouts (symmetric)
GRE1107	1107	5664	Problem from Grenoble collection
WEST2021	2021	7353	Chemical engineering: Column section

EDAGS outperforms the other two.

It is interesting to compare the performance on the three contrived examples. The analysis in the previous section can be used to show that Algorithm FILL1 takes $\Theta(n^3)$ time on EG-1 and $\Theta(n^2)$ time on EG-2 and EG-3; Algorithm FILL2 takes $\Theta(n^3)$ time on EG-2 and $\Theta(n^2)$ time on EG-1 and EG-3; and Algorithm EDAGS takes $\Theta(n^3)$ time on EG-3 and $\Theta(n^2)$ time on EG-1 and EG-2. This analysis is reflected in the measured times.

Table 3 compares the sizes of the factor matrices and their elimination dags, both determined by EDAGS. (In the table, $|G|$ means the number of edges in the graph G.) We see that in the worst-case example, EG-3, there is no difference, but in all the practical problems the edags are substantially smaller than the corresponding triangular factors. We also note that the symmetric matrix DWT0918, which is of order 918, has edags with 917 edges apiece; these are in fact both equal to its elimination tree.

Finally, for curiosity's sake, we present the result of one more experiment on the Harwell–Boeing matrices. Many of these matrices are highly reducible. Reducible linear systems are often solved by permuting to block triangular form and then factoring only the irreducible diagonal blocks. For each matrix in the test set, we first

$$\begin{pmatrix} 1 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & 2 & & & & & & \\ \bullet & & 3 & & & & & \\ \bullet & & & 4 & & & & \\ \bullet & & & & 5 & & & \\ \bullet & & & & & 6 & & \\ \bullet & & & & & & 7 & \\ \bullet & & & & & & & 8 \end{pmatrix} \quad \begin{pmatrix} 1 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & 2 & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ & \bullet & 3 & \bullet & \bullet & \bullet & \bullet & \bullet \\ & & \bullet & 4 & \bullet & \bullet & \bullet & \bullet \\ & & & \bullet & 5 & \bullet & \bullet & \bullet \\ & & & & \bullet & 6 & \bullet & \bullet \\ & & & & & \bullet & 7 & \bullet \\ & & & & & & \bullet & 8 \end{pmatrix} \quad \begin{pmatrix} 1 & & & & \bullet & \bullet & \bullet & \bullet \\ \bullet & 2 & & & & & & \\ \bullet & & 3 & & & & & \\ \bullet & & & 4 & & & & \\ & & & & 5 & & & \\ & & & & & 6 & & \\ & & & & & & 7 & \\ & & & & & & & 8 \end{pmatrix}$$

EG-1 (Arrow) EG-2 (Hessenberg) EG-3 (Worst Case)

FIG. 10. *Three contrived matrix examples.*

TABLE 2
Time to compute $G(L+U)$ from A.

Problem	n	Nonzeros	FILL1	FILL2	EDAGS
SHL400	663	1712	0.65	0.65	0.48
FS7603	760	5976	122.50	5.63	5.51
MCFE	765	24382	3.41	12.81	1.15
BP1600	822	4841	14.47	2.85	2.05
YOUNG3C	841	3988	1.05	4.49	0.60
DWT918	918	7384	25.59	7.69	2.84
GRE1107	1107	5664	14.89	6.93	4.34
WEST2021	2021	7353	39.80	14.93	7.66
EG-1	500	1498	68.69	1.64	3.12
EG-2	500	125749	0.60	68.78	2.12
EG-3	500	999	0.19	0.34	21.85

applied a row permutation to place nonzeros on the diagonal, then permuted the matrix to block upper triangular form, and then extracted the largest diagonal block B. Then we reordered that block by a symmetric minimum-degree permutation (on the structure of $B+B^T$). Table 4 presents the timings for the three fill algorithms on the resulting matrices. (There is no entry for SHL400, because that matrix is actually a permutation of a triangular matrix, so its largest irreducible diagonal block is 1 × 1.) We see that EDAGS is often fastest by a large margin, but FILL1 wins in some cases. The reason may be that these matrices, when reordered, have factors that are already so sparse that there is little to gain from transitive reduction.

6. Concluding remarks. This paper has defined a generalization of the elimination tree, which is an important structure in sparse Cholesky factorization, to an "elimination dag" useful for sparse unsymmetric LU factorization. The elimination dag arises by generalizing one particular property of the elimination tree, namely that it is the unique transitive reduction of the graph of the triangular factor. We have shown that elimination dags model the path structure of LU factors in a way that parallels, and generalizes, the elimination tree model of the Cholesky factor. We presented a new algorithm for sparse symbolic LU factorization, using edags, and showed experimentally that it compares favorably with existing symbolic schemes.

TABLE 3
Structural statistics for Algorithm EDAGS *in Fig.* 6.

Problem	n	Nonzeros	$\lvert G(L)\rvert$	$\lvert G^\circ(L)\rvert$	$\lvert G(U)\rvert$	$\lvert G^\circ(U)\rvert$
SHL400	663	1712	9191	892	8181	1355
FS7603	760	5976	200187	778	205817	779
MCFE	765	24382	23779	761	55911	761
BP1600	822	4841	66688	1459	68078	1431
YOUNG3C	841	3988	22066	845	22188	846
DWT918	918	7384	107347	917	107347	917
GRE1107	1107	5664	61087	3135	141887	1442
WEST2021	2021	7353	148368	10730	231052	4430
EG-1	500	1498	124750	499	124750	499
EG-1	500	125749	499	499	124750	499
EG-1	500	999	249	249	62500	62500

TABLE 4
Timings for largest irreducible blocks, ordered by minimum degree.

Problem	n	Nonzeros	FILL1	FILL2	EDAGS
FS7603 B	740	5878	0.80	1.43	0.40
MCFE B	697	23186	6.20	6.77	1.12
BP1600 B	217	1155	0.02	0.10	0.04
YOUNG3C B	841	3988	0.23	1.08	0.23
DWT918 B	918	7384	0.71	1.87	0.43
GRE1107 B	1107	5664	5.88	3.45	1.25
WEST2021 B	1500	5495	0.08	1.89	0.18

Eisenstat and Liu [5] have recently extended this work by defining a pair of dags they call the *symmetric reductions* of a matrix. These dags are intermediate in size between the edags and the triangular factors, but are easier to find than edags; Eisenstat and Liu's experiments show that a symbolic factorization algorithm based on them is even faster than our Algorithm EDAGS.

In this paper, we have used elimination dags only to study sparse LU factorization with no pivoting. When numerical pivoting is required for stability, we believe that edags can still be useful. The sparse partial pivoting scheme of Gilbert and Peierls [9] performs an LU factorization with pivoting in time proportional to arithmetic operations. Each column of the factors L and U is computed by a sparse triangular solve with the already-computed columns of L, of the form

$$\begin{pmatrix} L_j & 0 \\ L'_j & I \end{pmatrix}\begin{pmatrix} u_j \\ b_j \end{pmatrix} = a_j.$$

The triangular solve begins with a symbolic phase that uses the graph of the known part of L (actually, of L^T) to predict the nonzero structure of the result, in order to limit the numerical factorization to nonzero arithmetic. The symbolic phase for each column could be improved by using the elimination dag $G^\circ(L^T)$ instead of the graph $G(L^T)$. We expect this to be especially important when using vector machines, since the numerical phase can be speeded up by vectorization but the symbolic phase cannot. Eisenstat and Liu [4] have recently reported positive results using symmetric reductions in the symbolic phase of sparse partial pivoting on workstations.

REFERENCES

[1] A. V. Aho, M. R. Garey, and J. D. Ullman, *The transitive reduction of a directed graph*, SIAM J. Comput., 1 (1972), pp. 131–137.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, Reading, MA, 1983.

[3] I. S. Duff, R. Grimes, and J. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[4] S. C. Eisenstat and J. W. H. Liu, *Exploiting structural symmetry in a sparse partial pivoting code*, Presented at the Fourth SIAM Conf. Appl. Linear Algebra, Minneapolis, MN, 1991.

[5] ———, *Exploiting structural symmetry in unsymmetric sparse symbolic factorization*, Tech. Rep. CS–90–12, York University, North York, Ontario, Canada, 1990.

[6] A. George and J. W. H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[7] J. R. Gilbert, *Predicting structure in sparse matrix computations*, Tech. Rep. 86–750, Cornell University, Ithaca, NY, 1986; SIAM J. Matrix Anal. Appl., 15 (1994), to appear.

[8] J. R. Gilbert, C. Moler, and R. Schreiber, *Sparse matrices in Matlab: Design and implementation*, Tech. Rep. CSL 91–4, Xerox Palo Alto Research Center, Palo Alto, CA, 1991; SIAM J. Matrix Anal. Appl., 13 (1992), pp. 333–356.

[9] J. R. Gilbert and T. Peierls, *Sparse partial pivoting in time proportional to arithmetic operations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 862–874.

[10] J. W. H. Liu, *A compact row storage scheme for Cholesky factors using elimination trees*, ACM Trans. Math. Software, 12 (1986), pp. 127–148.

[11] ———, *The role of elimination trees in sparse factorization*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 134–172.

[12] D. J. Rose and R. E. Tarjan, *Algorithmic aspects of vertex elimination on directed graphs*, SIAM J. Appl. Math., 34 (1978), pp. 176–197.

[13] D. J. Rose, *Triangulated graphs and the elimination process*, J. Math. Anal. Appl., 32 (1970), pp. 597–609.

[14] R. Schreiber, *A new implementation of sparse Gaussian elimination*, ACM Trans. Math. Software, 8 (1982), pp. 256–276.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 353–371, April 1993

003

A SYSTOLIC ARRAY FOR SVD UPDATING*

MARC MOONEN[†], PAUL VAN DOOREN[‡], AND JOOS VANDEWALLE[†]

Abstract. In an earlier paper, an approximate SVD updating scheme has been derived as an interlacing of a QR updating on the one hand and a Jacobi-type SVD procedure on the other hand, possibly supplemented with a certain re-orthogonali zation scheme. This paper maps this updating algorithm onto a systolic array with $O(n^2)$ parallelism for $O(n^2)$ complexity, resulting in an $O(n^0)$ throughput. Furthermore, it is shown how a square root-free implementation is obtained by combining modified Givens rotations with approximate SVD schemes.

Key words. singular value decomposition, parallel algorithms, recursive least squares

AMS(MOS) subject classifications. 65F15, 65F25

CR classification. G.I.3

1. Introduction. The problem of continuously updating matrix decompositions as new rows are appended frequently occurs in signal processing applications. Typical examples are adaptive beamforming, direction finding, spectral analysis, pattern recognition, etc. [13].

In [12], it has been shown how an *SVD updating* algorithm can be derived by combining QR updating with a Jacobi-type SVD procedure applied to the triangular factor. In each time step an *approximate decomposition* is computed from a previous approximation at a *low computational cost*, namely, $O(n^2)$ operations. This algorithm was shown to be particularly suited for *subspace tracking* problems. The tracking error at each time step is then found to be bounded by the time variation in $O(n)$ time steps, which is sufficiently small for applications with slowly time-varying systems. Furthermore, the updating procedure was proved to be *stable* when supplemented with a certain re-orthogonalization scheme, which is elegantly combined with the updating.

In this paper, we show how this updating algorithm can be mapped onto a *systolic array* with $O(n^2)$ parallelism, resulting in an $O(n^0)$ throughput (similar to the case for mere QR updating; see [5]). Furthermore, it is shown how a square root-free implementation is obtained by combining modified Givens rotations with approximate SVD schemes.

In §2, the updating algorithm is briefly reviewed. A systolic implementation is described in §3 for the easy case, where corrective re-orthogonalizations are left out. In §4, it is shown how to incorporate these re-orthogonalizations. Finally, a square root-free implementation is derived in §5.

2. SVD updating. The *singular value decomposition* (SVD) of a real matrix $A_{m\times n}$ $(m \geq n)$ is a factorization of A into a product of three matrices

$$A_{m\times n} = U_{m\times n} \cdot \Sigma_{n\times n} \cdot V^T_{n\times n},$$

* Received by the editors November 10, 1989; accepted for publication (in revised form) October 18, 1991. This work was sponsored in part by BRA 3280 project of the European Community.

[†] ESAT, Katholieke Universiteit Leuven, K. Mercierlaan 94, 3001 Heverlee, Belgium (moonen@esat.kuleuven.ac.be and vandewalle@esat.kuleuven.ac.be). The first author is a senior research assistant with the Belgian N.F.W.O. (National Fund for Scientific Research).

[‡] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 1101 West Springfield Avenue, Urbana, Illinois 61801 (vandooren@uicsl.csl.uiuc.edu).

where U has orthonormal columns, V is an orthogonal matrix, and Σ is a diagonal matrix, with the singular values along the diagonal.

Given that we have the SVD of a matrix A, we may need to calculate the SVD of a matrix $\underline{A}$ that is obtained after appending a new row to A.

$$\underline{A} = \begin{bmatrix} A \\ a^T \end{bmatrix} = \underline{U}_{(m+1)\times n} \cdot \underline{\Sigma}_{n\times n} \cdot \underline{V}^T_{n\times n}.$$

In on-line applications, a new updating is often to be performed after each sampling. The data matrix at time step k is then defined in a recursive manner ($k \geq n$)

$$A^{(k)} = \begin{bmatrix} \lambda^{(k)} \cdot A^{(k-1)} \\ a^{(k)^T} \end{bmatrix} = U^{(k)}_{k\times n} \cdot \Sigma^{(k)}_{n\times n} \cdot V^{(k)^T}_{n\times n}.$$

Factor $\lambda^{(k)}$ is a *weighting factor*, and $a^{(k)}$ is the *measurement vector* at time instance k. For the sake of brevity, we consider only the case where $\lambda^{(k)}$ is a constant λ, although everything can easily be recast for the case where it is time varying. Finally, in most cases the $U^{(k)}$ matrices (of growing size!) need not be computed explicitly, and only $V^{(k)}$ and $\Sigma^{(k)}$ are explicitly updated.

An *adaptive algorithm* can be constructed by interlacing a Jacobi-type SVD procedure (Kogbetliantz's algorithm [9], modified for triangular matrices [8], [10]) with repeated QR updates. See [12] for further details.

```
Initialization
                    $V^{(0)} \Leftarrow I_{n\times n}$
                    $R^{(0)} \Leftarrow O_{n\times n}$

Loop
    for $k = 1, \ldots, \infty$
      input new measurement vector $a^{(k)}$

          $a_\star^{(k)^T} \Leftarrow a^{(k)^T} \cdot V^{(k-1)}$
          $R_\lambda^{(k)} \Leftarrow \lambda \cdot R^{(k-1)}$

      QR updating
          $\begin{bmatrix} R_\star^{(k)} \\ 0 \end{bmatrix} \Leftarrow Q^{(k)^T} \cdot \begin{bmatrix} R_\lambda^{(k)} \\ a_\star^{(k)^T} \end{bmatrix}$
          $R^{(k)} \Leftarrow R_\star^{(k)}$
          $V^{(k)} \Leftarrow V^{(k-1)}$

      SVD steps
        for $i = 1, \ldots, n-1$

            $R^{(k)} \Leftarrow \Theta_i^{(k)^T} \cdot R^{(k)} \cdot \Phi_i^{(k)}$
            $V^{(k)} \Leftarrow V^{(k)} \cdot \Phi_i^{(k)}$
            $\{V^{(k)} \Leftarrow T_i^{(k)} \cdot V^{(k)}\}$

        end
    end
```

Matrices $\Theta_i^{(k)}$ and $\Phi_i^{(k)}$ represent plane rotations (ith rotation in time step k) through angles $\theta_i^{(k)}$ and $\phi_i^{(k)}$ in the $(i, i+1)$-plane. The rotation angles $\theta_i^{(k)}$ and $\phi_i^{(k)}$ should be chosen such that the $(i, i+1)$ element in $R^{(k)}$ is zeroed, while $R^{(k)}$ remains in upper triangular form. Each iteration thus amounts to solving a 2×2 SVD on the main diagonal. The updating algorithm then reduces to applying sequences of $n-1$ rotations, where the *pivot index* i repeatedly takes up all the values $i = 1, 2, \ldots, n-1$ (n such sequences constitute a pipelined double sweep [14]), interlaced with QR updates. At each time step, $R^{(k)}$ will be "close" to a (block) diagonal matrix, so that in some sense $V^{(k)}$ is "close" to the exact matrix with right singular vectors; see [12].

Transformations $T_i^{(k)}$ correspond to approximate re-orthogonalizations of row vectors of $V^{(k)}$. These should be included in order to avoid round-off error buildup, if the algorithm is supposed to run for, say, thousands of time steps (see [12]). For the sake of clarity, the re-orthogonalizations are left out for a while, and are dealt with only in §4.

In the sequel, the time index k is often dropped for the sake of conciseness.

3. A systolic array for SVD updating. The above SVD updating algorithm—for the time being without re-orthogonalizations—can be mapped elegantly onto a systolic array, by combining systolic implementations for the matrix-vector product, the QR updating, and the SVD. In particular, with $n-1$ SVD iterations after each QR update,[1] an efficient parallel implementation is conceivable with $O(n^2)$ parallelism for $O(n^2)$ complexity. The SVD updating is then performed at a speed comparable to the speed of merely QR updating.

The SVD updating array is similar to the triangular SVD array in [10], where the SVD diagonalization and a preliminary QR factorization are performed on the same array. As for the SVD updating algorithm, the diagonalization process and the QR updating are interlaced, so that the array must be modified accordingly. Also, from the algorithmic description, it follows that the V-matrix should be stored as well. Hence, we have to provide for an additional square array, which furthermore performs the matrix-vector products $a^T \cdot V$. It is shown how the the matrix-vector product, the QR updating, and the SVD can be pipelined perfectly at the cost of little computational overhead. Finally, it is briefly shown how, e.g., a total least squares solution can be generated at each time step, with only very few additional computations.

Figure 1 gives an overview of the array. New data vectors are continuously fed into the left-hand side of the array. The matrix-vector product is computed in the square part, and the resulting vector is passed on to the triangular array that performs the QR updating and the SVD diagonalization. Output vectors are flushed upwards in the triangular array and become available at the right-hand side of the square array. All these operations can be carried out simultaneously, as is detailed next. The correctness of the array has also been verified by software simulation.

We first briefly review the SVD array of [10], and then modify the Gentleman–Kung QR updating array accordingly. Next, we show how to interlace the matrix-vector products, the QR updates, and the SVD process, and additionally generate (total least squares) output vectors.

3.1. SVD array. Figure 2 shows the SVD array of [10]. Processors on the main diagonal perform 2×2 SVDs, annihilating the available off-diagonal elements. Row

[1] If the number of rotations after each QR update is, for instance, halved or doubled, the array can easily be modified accordingly.

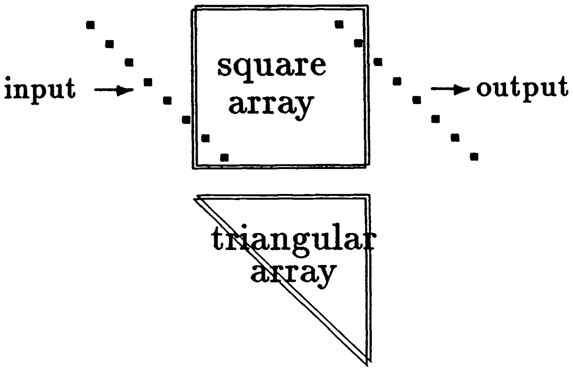


FIG. 1. *Overview* SVD *updating array.*

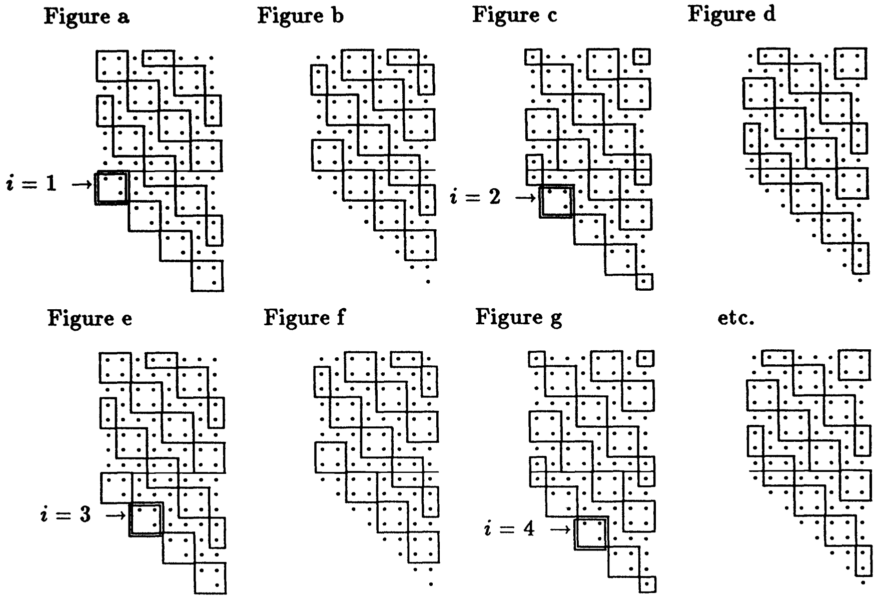


FIG. 2. SVD *array.*

transformation parameters are passed on to the right, while column transformation parameters are passed on upwards. Off-diagonal processors only apply and propagate these transformations to the next blocks outward. Column transformations are also propagated through the upper square part, containing the V-matrix (V's first row in the top row, etc.).

In this parallel implementation, off-diagonal elements with odd and even row numbers are being zeroed in an alternating fashion (*odd–even ordering*). However, it can easily be verified that an odd–even ordering corresponds to a cyclic-by-row or -column ordering, apart from a different start-up phase [11], [14]. The 2×2 SVDs that are performed in parallel on the main diagonal can indeed be thought of

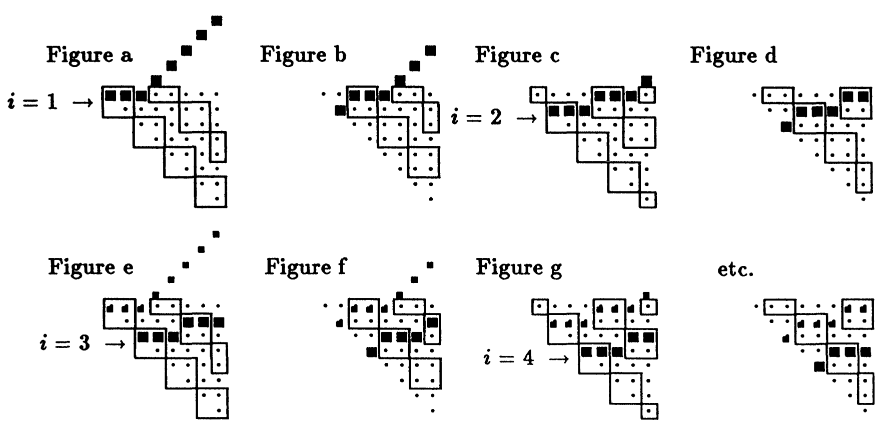


FIG. 3. *Modified Gentleman–Kung array.*

as corresponding to different pipelined sequences of $n-1$ rotations, where in each sequence the pivot index successively takes up the values $i = 1, \ldots, n-1$. A series of n such sequences is known to correspond to a double sweep (pipelined forward + backward) in a cyclic-by-rows ordering. In Fig. 2, one such sequence is indicated with double frames (for $i = 1, \ldots, 7$), starting in Fig. 2(a). In a similar fashion, the next sequence starts off from the top left corner in Fig. 2(e). As pointed out in §2, the QR updatings should be inserted in between two such sequences.

3.2. A modified Gentleman–Kung QR updating array. A QR updating is performed by applying a sequence of orthogonal transformations (*Givens rotations*) [6]. Gentleman and Kung have shown how pipelined sequences of Givens rotations can be implemented on a systolic array (see [5]). This array should now be matched to the SVD array, such that both can be combined.

Figure 3 shows a modified QR updating array. While all operations remain unaltered, the pipelining is somewhat different, so that the data vectors are now propagated through the array in a slightly different manner. The data vectors are fed into the array in a skewed fashion, as indicated, and are propagated downwards while being changed by successive row transformations. On the main diagonal, elementary orthogonal row transformations are generated. Rotation parameters are propagated to the right, while the transformed data vector components are passed on downwards. Note that each 2×2 block combines its first row with the available data vector components and pushes the resulting data vector components one step downwards. The first update starts off in Fig. 3(a) (large, filled boxes), the second in Fig. 3(e) (smaller, filled boxes), etc. Furthermore, each update is seen to correspond to a sequence of rotations where the pivot index takes up the values $i = 1, \ldots, n$. Both the processor's configuration and the pipelining turn out to be the same as for the SVD array.

3.3. Matrix-vector product. The matrix-vector product $a^T \cdot V$ can be combined with the SVD steps, as depicted in Figs. 4(a)–(g). The data vectors a^T are fed into the array in a skewed fashion, as indicated, and are propagated to the right, *in between two rotation fronts corresponding to the* SVD *diagonalization* (*frames*). Each processor receives a-components from its left neighbor, and intermediate results from its lower neighbor. The intermediate results are then updated and passed on to the upper neighbor, while the a-component is passed on to the right. The resulting

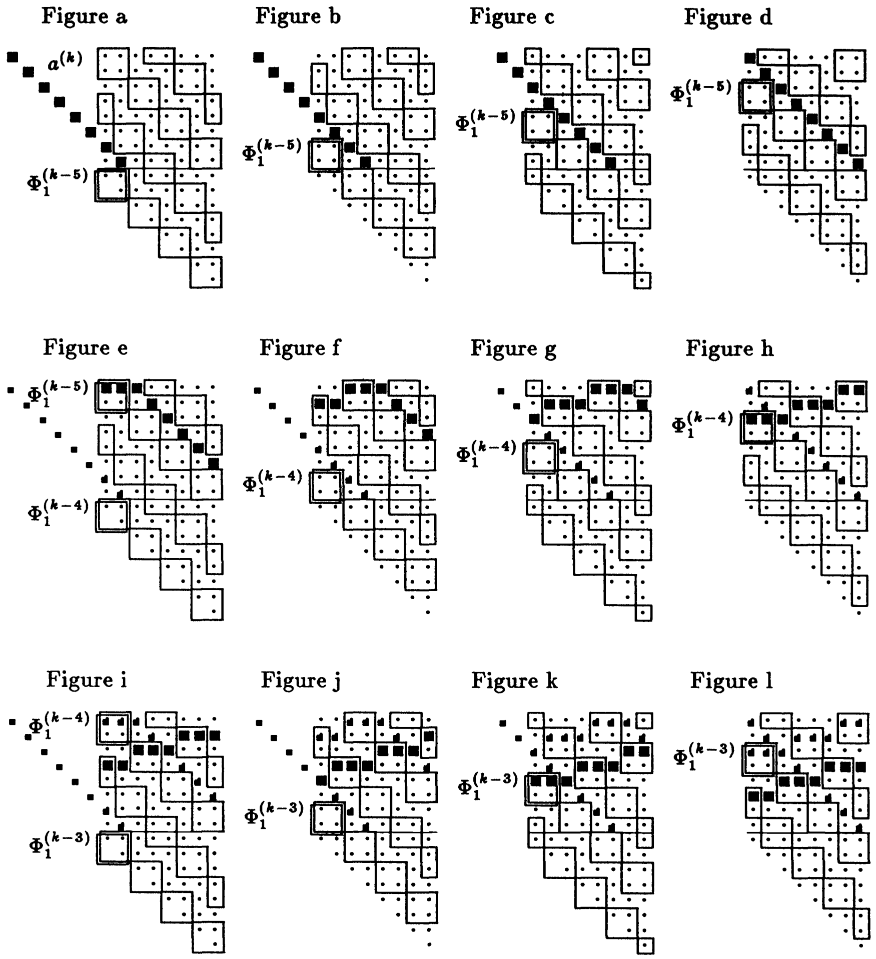


FIG. 4. SVD *updating array.*

matrix-vector product becomes available at the top end of the square array.

It should be stressed that a consistent matrix-vector product $a \cdot V$ *can only* be formed in between two SVD rotation fronts. That is a restriction, and it is worthwhile analyzing its implications.

—First, the propagation of the SVD rotation fronts dictates the direction in which a matrix-vector product can be formed. The resulting vector $a_\star$ thus inevitably becomes available at the top end of the square array, while it should be fed into the triangular array at the bottom for the subsequent QR update. The $a_\star$-components therefore have to be reflected at the top end and propagated downwards, towards the triangular array (Figs. 4(e)–(p)). The downward propagation of an $a_\star$-vector is then carried out in exactly the same manner as the propagation in the modified Gentleman–Kung array (see also Fig. 3).

—Second, the V-matrix that is used for computing $a^T \cdot V$ is in fact some older

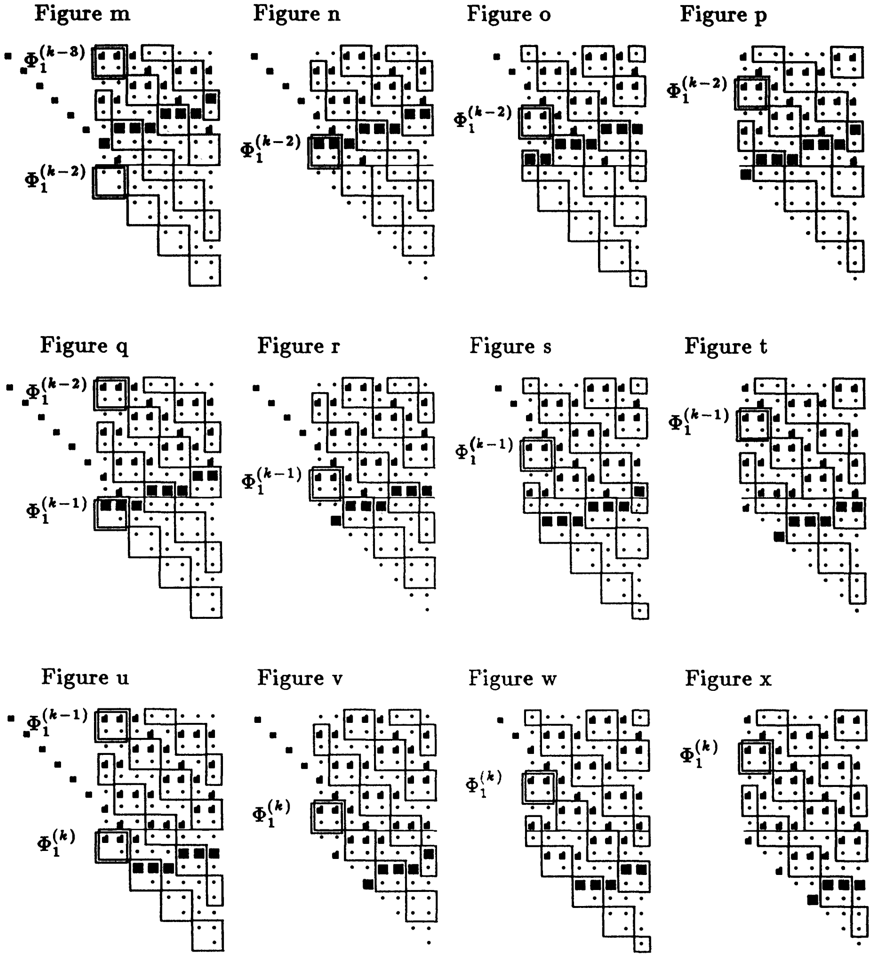


FIG. 4 (*continued*).

version of V, which we term $V^{(\natural)}$.[2] For a specific input vector $a^{(k)}$, this $V^{(\natural)}$ equals $V^{(k-1)}$ up to a number of column transformations, such that

$$V^{(k-1)} = V^{(\natural)} \cdot \Phi^{(\natural)},$$

where $\Phi^{(\natural)}$ denotes the accumulated column transformations. In order to obtain $a_\star^{(k)}$, it is necessary to apply $\Phi^{(\natural)}$ to the computed matrix-vector product

$$a_\star^{(k)} = a^{(k)^T} \cdot V^{(k-1)} = \underbrace{a^{(k)^T} \cdot V^{(\natural)}}_{a_\star^{(\natural)^T}} . \Phi^{(\natural)}.$$

These additional transformations represent a computational overhead, which is the penalty for pipelining the matrix-vector products with the SVD steps on the same

[2] One can check that it is not possible to substitute a specific time index for the "$\natural$."

array. Notice, however, that at the same time, the throughput improves greatly. Waiting until $V^{(k-1)}$ is formed completely before calculating the matrix-vector product would induce $O(n)$ time lags and likewise result in an $O(n^{-1})$ throughput. With the additional computations, the throughput is $O(n^0)$.

Let us now focus on the transformations in $\Phi^{(\natural)}$ and the way these can be processed. One of these transformations is, e.g., $\Phi_1^{(k-5)}$ (see §2 for notation), which is computed on the main diagonal in Fig. 4(a) (double frame). While propagating downwards, the ${a_\star^{(\natural)}}^T$-vector crosses the upgoing rotation $\Phi_1^{(k-5)}$ in Fig. 4(e). At this point, this transformation can straightforwardly be applied to the available ${a_\star^{(\natural)}}^T$-components. Similarly, one can verify that $\Phi_1^{(k-4)}$, $\Phi_1^{(k-3)}$, $\Phi_1^{(k-2)}$, and $\Phi_1^{(k-1)}$ are applied in Figs. 4(h), 4(k), 4(n), and 4(q), respectively. The transformations in the other columns can be traced similarly. In conclusion, each frame in the square array now corresponds to a column transformation that is applied to a 2×2 block of the V-matrix *and* to the two available components of an $a_\star^{(\natural)}$-vector. These components are propagated one step downwards next. A complete description for a 2×2 block in the V-matrix is presented in Display 1. Notation is slightly modified for conciseness, and $\circ$ and $\diamond$ represent memory cells that are filled by the updated elements of the $a_\star^{(\natural)}$-vector.

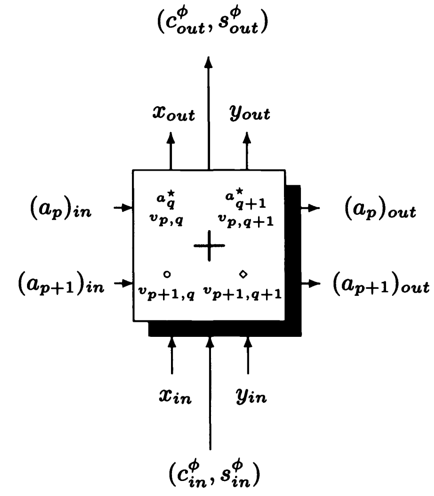
By the end of Fig. 4(p), the first $a_\star$-vector leaves the square array in a form directly amenable to the (modified Gentleman–Kung) triangular array.

3.4. Interlaced QR updating and SVD diagonalization. Finally, the modified Gentleman–Kung array and the triangular SVD array are easily combined (Fig. 4(q)–(x)). In each frame, column and row transformations corresponding to the SVD diagonalization are performed first (see also Fig. 2), while in a second step, only row transformations are performed corresponding to the modified QR updating (affecting the $a_\star$-components and the upper part of the 2×2 -blocks (see also Fig. 3). Again, column transformations in the first step should be applied to the $a_\star$-components as well. Boundary cells and internal cells are described in Displays 2 and 3.

Without disturbing the array operations, it is possible to output particular singular vectors (e.g., total least squares solutions [15]) at regular time intervals. This is easily done by performing matrix-vector multiplications $V \cdot t$, where t is a vector with all its components equal to zero, except for one component equal to 1, and which is generated on the main diagonal. The t-vector is propagated upwards to the square array, where the matrix-vector product $V \cdot t$ is performed, which singles out the appropriate right singular vector. While t is propagated upwards, intermediate results are propagated to the right, such that the resulting vector becomes available at the right-hand side of the array. These solution vectors can be generated at the same rate as the input data vectors are fed in, and both processes can run simultaneously without interference.

4. Including re-orthogonalizations.

In [12], it was shown how additional re-orthogonalizations stabilize the overall round-off error propagation in the updating scheme. In the algorithmic description of §2, $T_i^{(k)}$ is an approximate re-orthogonalization and normalization of rows p and q in the V-matrix. The row indices p and q are chosen as functions of k and i, in a cyclic manner. Furthermore, the re-orthogonalization scheme was shown to converge quadratically. In view of efficient parallel implementation, we first reorganize this re-orthogonalization scheme. The modified scheme is then easily mapped onto the systolic array. The computational overhead



Input Transformation Parameters

$$c^{\phi}, s^{\phi} \leftarrow c^{\phi}_{in}, s^{\phi}_{in}$$

Apply Transformation

$$\begin{bmatrix} a^{\star}_q & a^{\star}_{q+1} \\ v_{p,q} & v_{p,q+1} \\ v_{p+1,q} & v_{p+1,q+1} \end{bmatrix}$$

$$\leftarrow \begin{bmatrix} a^{\star}_q & a^{\star}_{q+1} \\ v_{p,q} & v_{p,q+1} \\ v_{p+1,q} & v_{p+1,q+1} \end{bmatrix} \begin{bmatrix} s^{\phi} & c^{\phi} \\ c^{\phi} & -s^{\phi} \end{bmatrix}$$

Propagate $a^{\star}_q, a^{\star}_{q+1}$

$$\circ, \diamond \leftarrow a^{\star}_q, a^{\star}_{q+1}$$

Propagate Transformation Parameters

$$c^{\phi}_{out}, s^{\phi}_{out} \leftarrow c^{\phi}, s^{\phi}$$

if $q =$ **even**

Input a_p, a_{p+1} and Intermediate Results

$$a_p, a_{p+1}, x, y \leftarrow (a_p)_{in}, (a_{p+1})_{in}, x_{in}, y_{in}$$

Update Intermediate Result

$$\begin{aligned} x &\leftarrow x + v_{p,q} \cdot a_p + v_{p+1,q} \cdot a_{p+1} \\ y &\leftarrow y + v_{p,q+1} \cdot a_p + v_{p+1,q+1} \cdot a_{p+1} \end{aligned}$$

Propagate a_p, a_{p+1} and intermediate results

$$(a_p)_{out}, (a_{p+1})_{out}, x_{out}, y_{out} \leftarrow a_p, a_{p+1}, x, y$$

end

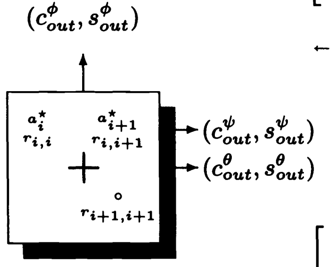
DISPLAY 1. *Internal cell V-matrix.*

turns out to be negligible, as the square part of the array (V-matrix) so far remained underloaded, compared to the triangular part (see below for figures).

First of all, as the re-orthogonalization scheme cyclicly adjusts the row vectors in the V-matrix, it is straightforward to introduce additional *row permutations* in the square part of the array. The 2×2 blocks in the square part then correspond to column transformations (SVD scheme) and row permutations (re-orthogonalization scheme). Orthogonal column transformations clearly do not affect the norms and inner products of the rows, except for local rounding errors assumed smaller than the accumulated errors. Hence, the column transformations are assumed not to interfere with the re-orthogonalization and thus need not be considered anymore. As for the row permutations, subsequent positions for the elements in the first column of V are indicated in Fig. 5, for a (fairly) arbitrary initial row numbering (as an example, the 2×2 block in the upper-left corner in Fig. 5(a) interchanges elements 4 and 5, etc.).

Let us now focus on *one* single row (row 1) and see how it can (approximately) be normalized and orthogonalized with respect to *all other* rows. Later, we will use this in an overall procedure.

1. In a first step, the norm (squared) and inner products are computed as a matrix-vector product


$(c^{\phi}_{out}, s^{\phi}_{out})$
$a^{\star}_{i}$ $a^{\star}_{i+1}$
$r_{i,i}$ $r_{i,i+1}$
+
o
$r_{i+1,i+1}$
$(c^{\psi}_{out}, s^{\psi}_{out})$
$(c^{\theta}_{out}, s^{\theta}_{out})$

Compute Rotation Angles ϕ, θ

$$c^{\phi}, s^{\phi}, c^{\theta}, s^{\theta} \leftarrow \cos\phi, \sin\phi, \cos\theta, \sin\theta$$

Apply Transformation

$$\begin{bmatrix} a^{\star}_{i} & a^{\star}_{i+1} \\ r_{i,i} & 0 \\ & r_{i+1,i+1} \end{bmatrix} \leftarrow \begin{bmatrix} 1 & & \\ & s^{\theta} & c^{\theta} \\ & c^{\theta} & -s^{\theta} \end{bmatrix} \begin{bmatrix} a^{\star}_{i} & a^{\star}_{i+1} \\ r_{i,i} & r_{i,i+1} \\ & r_{i+1,i+1} \end{bmatrix} \begin{bmatrix} s^{\phi} & c^{\phi} \\ c^{\phi} & -s^{\phi} \end{bmatrix}$$

Compute Rotation Angle ψ

$$c^{\psi}, s^{\psi} \leftarrow \cos\psi, \sin\psi$$

Apply Transformation

$$\begin{bmatrix} r_{i,i} & r_{i,i+1} \\ 0 & a^{\star}_{i+1} \end{bmatrix} \leftarrow \begin{bmatrix} c^{\psi} & s^{\psi} \\ -s^{\psi} & c^{\psi} \end{bmatrix} \begin{bmatrix} r_{i,i} & 0 \\ a^{\star}_{i} & a^{\star}_{i+1} \end{bmatrix}$$

Propagate $a^{\star}_{i+1}$

$$\circ \leftarrow a^{\star}_{i+1}$$

Propagate Transformation Parameters

$$c^{\phi}_{out}, s^{\phi}_{out}, c^{\theta}_{out}, s^{\theta}_{out}, c^{\psi}_{out}, s^{\psi}_{out} \leftarrow c^{\phi}, s^{\phi}, c^{\theta}, s^{\theta}, c^{\psi}, s^{\psi}$$

DISPLAY 2. *Boundary cell R-factor.*

$$x_1 \overset{\text{def}}{=} V \cdot v_1 = \begin{bmatrix} \boxed{v_1^T} \\ \boxed{v_2^T} \\ \vdots \\ \boxed{v_n^T} \end{bmatrix} \cdot \begin{bmatrix} \boxed{v_1} \end{bmatrix} \overset{\text{def}}{=} \begin{bmatrix} \xi_{11} \\ \xi_{12} \\ \vdots \\ \xi_{1n} \end{bmatrix},$$

where v_i^T is the ith row in V. On the systolic array, where v_1 initially resides in the bottom row, it suffices to propagate the v_1-components upwards, and accumulate the inner products from the left to the right (Figs. 5(a)–(h), where pq is shorthand for ξ_{pq}). The resulting x_1-vector components run out at the right-hand side.

2. In a second step, this x_1-vector is back-propagated to the left (Figs 5(e)–(t)). Due to the permutations along the way, the x_1-components reach the left-hand side of the array at the right time and in the right place, such that

Input Transformation Parameters

$$c^{\phi}, s^{\phi}, c^{\theta}, s^{\theta}, c^{\psi}, s^{\psi} \leftarrow c^{\phi}_{in}, s^{\phi}_{in}, c^{\theta}_{in}, s^{\theta}_{in}, c^{\psi}_{in}, s^{\psi}_{in}$$

Apply Transformation

$$\begin{bmatrix} a^{\star}_{q} & a^{\star}_{q+1} \\ r_{p,q} & r_{p,q+1} \\ r_{p+1,q} & r_{p+1,q+1} \end{bmatrix} \leftarrow \begin{bmatrix} 1 & & \\ & s^{\theta} & c^{\theta} \\ & c^{\theta} & -s^{\theta} \end{bmatrix} \begin{bmatrix} a^{\star}_{q} & a^{\star}_{q+1} \\ r_{p,q} & r_{p,q+1} \\ r_{p+1,q} & r_{p+1,q+1} \end{bmatrix} \begin{bmatrix} s^{\phi} & c^{\phi} \\ c^{\phi} & -s^{\phi} \end{bmatrix}$$

Apply Transformation

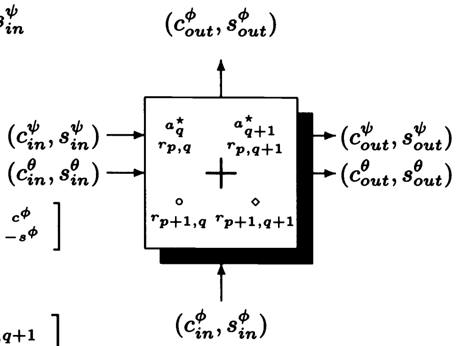
$$\begin{bmatrix} r_{p,q} & r_{p,q+1} \\ a^{\star}_{q} & a^{\star}_{q+1} \end{bmatrix} \leftarrow \begin{bmatrix} c^{\psi} & s^{\psi} \\ -s^{\psi} & c^{\psi} \end{bmatrix} \begin{bmatrix} r_{p,q} & r_{p,q+1} \\ a^{\star}_{q} & a^{\star}_{q+1} \end{bmatrix}$$

Propagate $a^{\star}_{q}, a^{\star}_{q+1}$

$$\circ, \diamond \leftarrow a^{\star}_{q}, a^{\star}_{q+1}$$

Propagate Transformation Parameters

$$c^{\phi}_{out}, s^{\phi}_{out}, c^{\theta}_{out}, s^{\theta}_{out}, c^{\psi}_{out}, s^{\psi}_{out} \leftarrow c^{\phi}, s^{\phi}, c^{\theta}, s^{\theta}, c^{\psi}, s^{\psi}$$


$(c^{\phi}_{out}, s^{\phi}_{out})$
$(c^{\psi}_{in}, s^{\psi}_{in})$
$(c^{\theta}_{in}, s^{\theta}_{in})$
$a^{\star}_{q}$ $a^{\star}_{q+1}$
$r_{p,q}$ $r_{p,q+1}$
+
$\circ$ $\diamond$
$r_{p+1,q}$ $r_{p+1,q+1}$
$(c^{\psi}_{out}, s^{\psi}_{out})$
$(c^{\theta}_{out}, s^{\theta}_{out})$
$(c^{\phi}_{in}, s^{\phi}_{in})$

DISPLAY 3. *Internal cell R-factor.*

3. in a third step, a *correction vector* can be computed as a matrix-vector product

$$y_1 \overset{\text{def}}{=} \begin{bmatrix} \frac{1}{2}(\xi_{11} - 1) & \xi_{12} & \cdots & \xi_{1n} \end{bmatrix} \cdot \begin{bmatrix} \boxed{v_1^T} \\ \boxed{v_2^T} \\ \vdots \\ \boxed{v_n^T} \end{bmatrix},$$

where the term $\frac{1}{2}(\xi_{11} - 1)$ corresponds to the first-order term in the Taylor series expansion for the normalization of v_1. The y_1-vector components are accumulated from the bottom to the top, while x_1 is again being propagated to the right (Fig. 5(q)–(w)). Finally,

4. in a fourth step, v_1, which meanwhile moved on to the top row of the array, is adjusted with y_1:

$$v_1^{\star} \leftarrow v_1 - y_1.$$

These operations are performed in the top row of the array (Figs. 5(t)–(w)). One can check that if

$$v_1 \cdot v_1 = 1 + O(\epsilon),$$
$$v_1 \cdot v_p = O(\epsilon), \qquad p = 2, \ldots, n$$

for some small $\epsilon \ll 1$, then

$$v_1^{\star} \cdot v_1^{\star} = 1 + O(\epsilon^2),$$

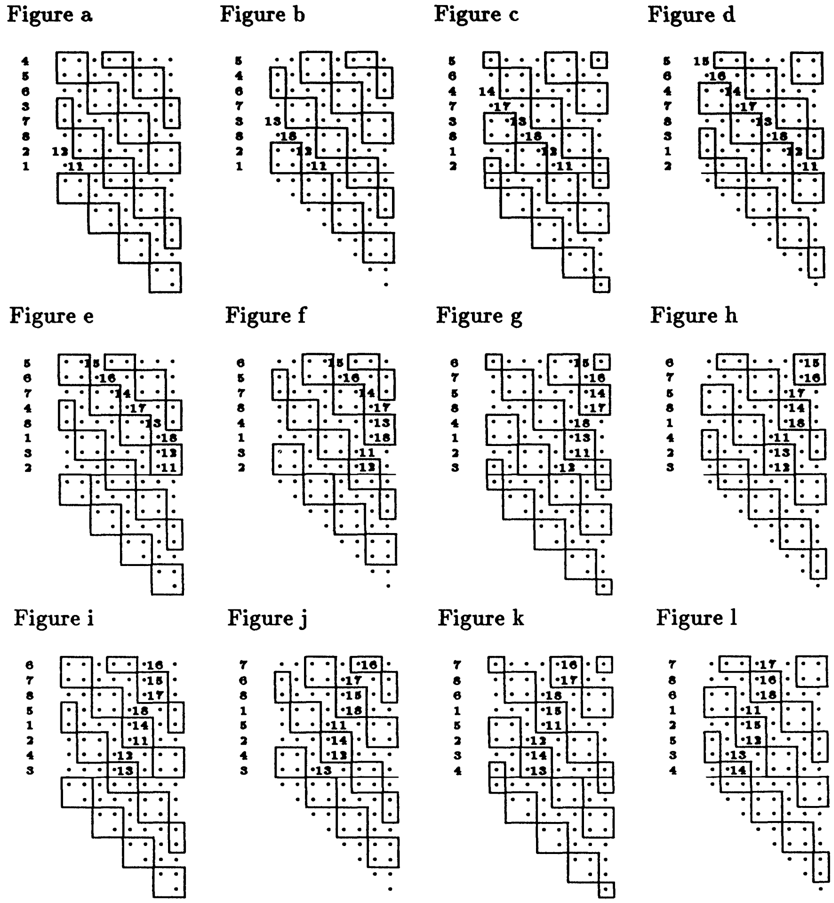


Figure a

Figure b

Figure c

Figure d

Figure e

Figure f

Figure g

Figure h

Figure i

Figure j

Figure k

Figure l

FIG. 5. *Re-orthogonalizations.*

$$v_1^\star \cdot v_p = O(\epsilon^2), \qquad p = 2, \dots, n.$$

The above procedure for v_1 should now be repeated for rows 2, 3, etc., and furthermore, everything should be pipelined. Obviously, one could start a similar procedure for v_2 in Fig. 5(e), for v_3 in Fig. 5(i), etc. The pipelining of such a scheme would be remarkably simple, but unfortunately there is something wrong with it. A slight modification is needed to make things work properly.

As for the processing of v_2, one easily checks that the computed inner product ξ_{21} equals $v_2 \cdot v_1 = v_1 \cdot v_2 = \xi_{12}$, while it *should* equal $v_2 \cdot v_1^\star$. A similar problem occurs with ξ_{31} and ξ_{32}, etc. In general, problems occur when computing inner products with *ascending rows*, which still have to be adjusted in the top row of the array before the relevant inner product can be computed. This problem is readily solved as follows. Instead of computing inner products with *all* other rows, we only take *descending rows* into account. This is easily done by assigning *tags* to the rows, where, e.g., a 0-tag indicates an ascending row, and a 1-tag indicates a descending row. Tags are

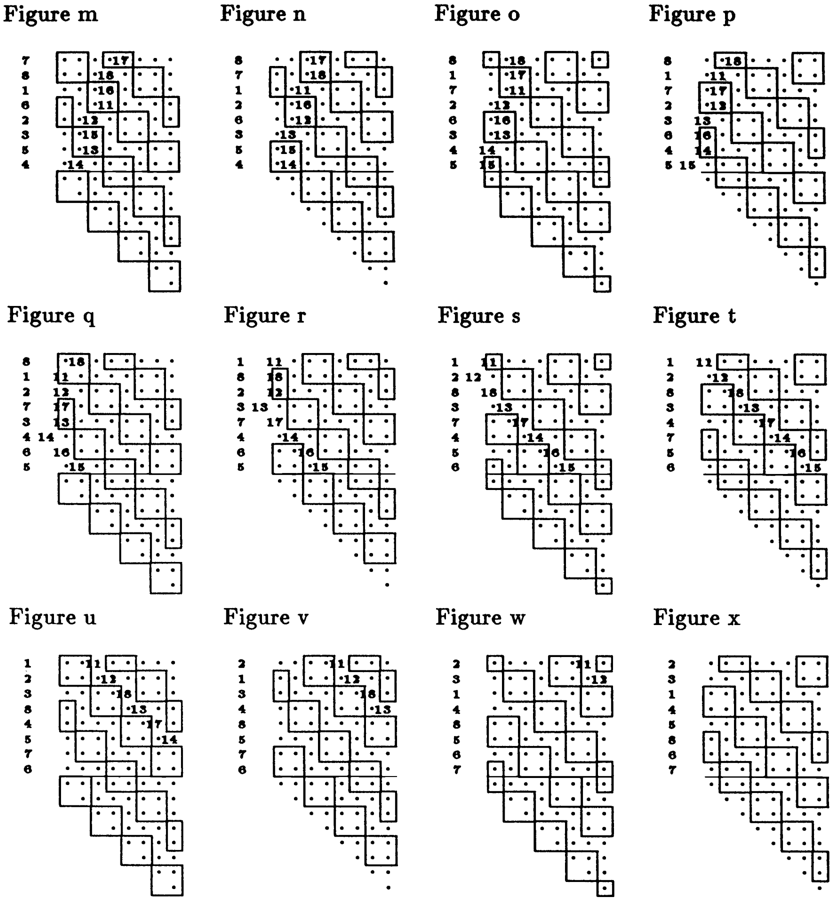


FIG. 5 (*continued*).

reset at the top and the bottom of the array. Computing an x_i vector is then done as follows:

$$x_i \overset{\text{def}}{=} V \cdot v_i = \begin{bmatrix} \boxed{\mathrm{TAG}_1 v_1^T} \\ \boxed{\mathrm{TAG}_2 v_2^T} \\ \vdots \\ \boxed{\mathrm{TAG}_n v_n^T} \end{bmatrix} \cdot \begin{bmatrix} \boxed{v_i} \end{bmatrix} \overset{\text{def}}{=} \begin{bmatrix} \xi_{i1} \\ \xi_{i2} \\ \vdots \\ \xi_{in} \end{bmatrix}.$$

For the 8 × 8 example of Fig. 5, one can check that the result of this is that v_1 is orthogonalized onto v_2, v_3, v_4, and v_5 ($\xi_{16} = \xi_{17} = \xi_{18} = 0$, because rows 6, 7, and 8 are ascending rows at that time, i.e., $\mathrm{TAG}_6 = \mathrm{TAG}_7 = \mathrm{TAG}_8 = 0$). Similarly, v_2 is orthogonalized onto v_3, v_4, v_5, and v_6 ($\xi_{27} = \xi_{28} = \xi_{21} = 0$), etc. The resulting ordering is recast as follows (pp refers to the normalization of row p, whereas pq for

TABLE 1

	SVD updating	Re-orthog.	Total
Internal cell V-matrix	16× 10+	8× 8+	24× 18+
Internal cell R-matrix	28× 14+		28× 14+
Diagonal cell R-matrix	25× 13+ 9÷ 4√		25× 13+ 9÷ 4√

$p \neq q$ refers to the orthogonalization of row p onto row q)

$$\begin{array}{cccccccccccc}
11 & 12 & 13 & 14 & 15 & & & & & & & \\
 & 22 & 23 & 24 & 25 & 26 & & & & & & \\
 & & 33 & 34 & 35 & 36 & 37 & & & & & \\
 & & & 44 & 45 & 46 & 47 & 48 & & & & \\
 & & & & 55 & 56 & 57 & 58 & 51 & & & \\
 & & & & & 66 & 67 & 68 & 61 & 62 & & \\
 & & & & & & 77 & 78 & 71 & 72 & 73 & \\
 & & & & & & & 88 & 81 & 82 & 83 & 84
\end{array}$$

In such a *sweep*, each combination appears at least once (wherewith $4 = \frac{n}{2}$ combinations appear twice, e.g., 15 (51), etc.). In the general case, the rows are cyclicly normalized and orthogonalized onto the $\frac{n}{2}$ succeeding rows. One can easily prove that if $\|VV^T - I\|_F = O(\epsilon)$ before a particular sweep, then $\|VV^T - I\|_F = O(\epsilon^2)$ after the sweep. In other words, the quadratic convergence rate is maintained. On the other hand, it is seen that one single sweep takes twice the computation time for a sweep in a "normal" cyclic by rows or odd–even ordering. This is hardly an objection, as the re-orthogonalization is only meant to keep V *reasonably close* to orthogonal. The error analysis in [12] thus still applies (with slightly adjusted constant coefficients).

Complete processor descriptions are left out for the sake of brevity. Let it suffice to give an operation count for different kinds of processors. See Table 1, from which it follows that the re-orthogonalizations do not increase the load of the critical cells.

Finally, as the rows of V continuously interchange, each input vector a in Fig. 1 should be permuted accordingly, before multiplication (see §2, $a^t \cdot V = (a^t \cdot P^t) \cdot (P \cdot V)$, where P is a permutation matrix). One can straightforwardly design a kind of "preprocessor" for a, which outputs the right components of a at the right time. For the sake of brevity, we will not go into details here.

5. Square root-free algorithms. The *throughput* in the parallel SVD updating array is essentially determined by the computation times for the processors on the main diagonal to calculate the rotation angles both for the QR updatings and the SVD steps. In general, these computations require, respectively, one and three square roots, which appears to be the main computational bottleneck. Gentleman developed a square root-free procedure for QR updating [1], [4], [7] where use is made of a (one-sided) factorization of the R-matrix. The SVD schemes as such, however, do not lend themselves to square root-free implementation. Still, in [3] a few alternative SVD schemes have been investigated based on *approximate* formulas for the computation of either $\tan\theta$ or $\tan\phi$. When combined with a (generalized) Gentleman procedure with a two-sided factorization of the R-factor, these schemes eventually yield square root-free SVD updating algorithms. Implementation on a systolic array hardly imposes any changes when compared to the conventional algorithm. Furthermore, as the approximate formulas for the rotation angles are in fact (at least) first-order approximations, the (first-order) performance analysis in [12] still applies. In other words, the same upper bounds for the tracking error are valid, even when approximate formulas are used.

5.1. Square root-free SVD computations. The SVD procedure is seen to reduce to solving elementary 2×2 SVDs on the main diagonal (§2). For an *approximate* SVD computation, the relevant transformation formula becomes

$$\begin{bmatrix} r^*_{i,i} & r^*_{i,i+1} \\ 0 & r^*_{i+1,i+1} \end{bmatrix} = \begin{bmatrix} -\sin\theta & \cos\theta \\ \cos\theta & \sin\theta \end{bmatrix} \begin{bmatrix} r_{i,i} & r_{i,i+1} \\ 0 & r_{i+1,i+1} \end{bmatrix} \begin{bmatrix} -\sin\phi & \cos\phi \\ \cos\phi & \sin\phi \end{bmatrix},$$

where $r_{i,i+1}$ is only being approximately annihilated $(|r^*_{i,i+1}| \le |r_{i,i+1}|)$. In particular, the following approximate schemes from [3] turn out to be very useful for our purpose. (For details, refer to [3].)

if $|r_{i,i}| \ge |r_{i+1,i+1}|$

$$\sigma = \frac{r_{i+1,i+1} r_{i,i+1}}{r^2_{i,i} - r^2_{i+1,i+1} + r^2_{i,i+1}}$$

approximation 1: $\tan\theta = \sigma$
approximation 2: $\tan\theta = \frac{\sigma}{1+\sigma^2}$

$$\tan\phi = \frac{r_{i+1,i+1}\tan\theta + r_{i,i+1}}{r_{i,i}}$$

if $|r_{i,i}| \le |r_{i+1,i+1}|$

$$\sigma = \frac{r_{i,i} r_{i,i+1}}{r^2_{i+1,i+1} - r^2_{i,i} + r^2_{i,i+1}}$$

approximation 1: $\tan\phi = \sigma$
approximation 2: $\tan\phi = \frac{\sigma}{1+\sigma^2}$

$$\tan\theta = \frac{r_{i,i}\tan\phi - r_{i,i+1}}{r_{i+1,i+1}}$$

In the sequel, we consider only $|r_{i,i}| \ge |r_{i+1,i+1}|$, as the derived formulas can straightforwardly be adapted for the other case. These approximate schemes still require two square roots for the computation of $\cos\phi$ and $\cos\theta$.

The above approximate formulas can, however, be combined with a (generalized) Gentleman procedure, where use is made of a two-sided factorization of the R-matrix

$$R = D^{\frac{1}{2}}_{row} \cdot \bar{R} \cdot D^{\frac{1}{2}}_{col}$$

and where only $\bar{R}$, D_{row}, and D_{col} are stored (in the sequel, an overbar always refers to a factorization).

Let us first rewrite the first approximate formula for $\tan\theta$:

$$\tan\theta = \underbrace{\left(\frac{r^2_{i+1,i+1}}{r^2_{i,i} - r^2_{i+1,i+1} + r^2_{i,i+1}}\right)}_{\rho} \cdot \frac{r_{i,i+1}}{r_{i+1,i+1}}.$$

As ρ contains only squared values, it can be computed from the factorization of R as well:

$$\rho = \frac{d^{row}_{i+1} d^{col}_{i+1} \bar{r}^2_{i+1,i+1}}{d^{row}_i d^{col}_i \bar{r}^2_{i,i} - d^{row}_{i+1} d^{col}_{i+1} \bar{r}^2_{i+1,i+1} + d^{row}_i d^{col}_{i+1} \bar{r}^2_{i,i+1}}.$$

Obviously, a similar formula for ρ can be derived from the second approximate formula for $\tan\theta$ (which has better convergence properties; see [3]). Applying Gentleman's procedure to the *row transformation* then gives

$$\begin{bmatrix} -\sin\theta & \cos\theta \\ \cos\theta & \sin\theta \end{bmatrix} \cdot \begin{bmatrix} \sqrt{d^{row}_i} & 0 \\ 0 & \sqrt{d^{row}_{i+1}} \end{bmatrix} \cdot \begin{bmatrix} \bar{r}_{i,i} & \bar{r}_{i,i+1} \\ 0 & \bar{r}_{i+1,i+1} \end{bmatrix} \cdot \begin{bmatrix} \sqrt{d^{col}_i} & 0 \\ 0 & \sqrt{d^{col}_{i+1}} \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{d^{row}_{i+1}}\cos\theta & 0 \\ 0 & \sqrt{d^{row}_i}\cos\theta \end{bmatrix} \cdot \begin{bmatrix} -\tan\theta\sqrt{\frac{d^{row}_i}{d^{row}_{i+1}}} & 1 \\ 1 & \tan\theta\sqrt{\frac{d^{row}_{i+1}}{d^{row}_i}} \end{bmatrix}$$

$$\cdot \begin{bmatrix} \bar{r}_{i,i} & \bar{r}_{i,i+1} \\ 0 & \bar{r}_{i+1,i+1} \end{bmatrix} \cdot \begin{bmatrix} \sqrt{d^{col}_i} & 0 \\ 0 & \sqrt{d^{col}_{i+1}} \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{d^{row^*}_i} & 0 \\ 0 & \sqrt{d^{row^*}_{i+1}} \end{bmatrix} \cdot \begin{bmatrix} \bar{r}^{\circ}_{i,i} & \bar{r}^{\circ}_{i,i+1} \\ \bar{r}^{\circ}_{i+1,i} & \bar{r}^{\circ}_{i+1,i+1} \end{bmatrix} \cdot \begin{bmatrix} \sqrt{d^{col}_i} & 0 \\ 0 & \sqrt{d^{col}_{i+1}} \end{bmatrix}.$$

With

$$\tan\theta = \rho \cdot \frac{\bar{r}_{i,i+1}\sqrt{d^{row}_i}}{\bar{r}_{i+1,i+1}\sqrt{d^{row}_{i+1}}},$$

this leads to row transformation formulas

$$\begin{bmatrix} \bar{r}^{\circ}_{i,i} & \bar{r}^{\circ}_{i,i+1} \\ \bar{r}^{\circ}_{i+1,i} & \bar{r}^{\circ}_{i+1,i+1} \end{bmatrix} = \begin{bmatrix} -\rho \cdot \frac{\bar{r}_{i,i+1} d^{row}_i}{\bar{r}_{i+1,i+1} d^{row}_{i+1}} & 1 \\ 1 & \rho \cdot \frac{\bar{r}_{i,i+1}}{\bar{r}_{i+1,i+1}} \end{bmatrix} \cdot \begin{bmatrix} \bar{r}_{i,i} & \bar{r}_{i,i+1} \\ 0 & \bar{r}_{i+1,i+1} \end{bmatrix}$$

with scale factor updating (notice the implicit row permutation)

$$d^{row^*}_i = d^{row}_{i+1}\cos^2\theta,$$
$$d^{row^*}_{i+1} = d^{row}_i\cos^2\theta,$$

$$\cos^2\theta = \frac{1}{1 + \rho^2 \cdot \frac{\bar{r}^2_{i,i+1} d^{row}_i}{\bar{r}^2_{i+1,i+1} d^{row}_{i+1}}}.$$

Note that due to the 1's in the first transformation formula, a 50 percent saving in the number of multiplications is obtained in the off-diagonal processors.

The *column transformation* should then annihilate $r^{\circ}_{i+1,i}$ in order to preserve the triangular structure. With

$$\tan\phi = \frac{\bar{r}^{\circ}_{i+1,i+1}\sqrt{d^{col}_{i+1}}}{\bar{r}^{\circ}_{i+1,i}\sqrt{d^{col}_{i}}}$$

we can again apply Gentleman's procedure as follows:

$$\begin{aligned}
&\begin{bmatrix} \sqrt{d^{row*}_{i}} & 0 \\ 0 & \sqrt{d^{row*}_{i+1}} \end{bmatrix} \cdot \begin{bmatrix} \bar{r}^{\circ}_{i,i} & \bar{r}^{\circ}_{i,i+1} \\ \bar{r}^{\circ}_{i+1,i} & \bar{r}^{\circ}_{i+1,i+1} \end{bmatrix} \cdot \begin{bmatrix} \sqrt{d^{col}_{i}} & 0 \\ 0 & \sqrt{d^{col}_{i+1}} \end{bmatrix} \cdot \begin{bmatrix} -\sin\phi & \cos\phi \\ \cos\phi & \sin\phi \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{d^{row*}_{i}} & 0 \\ 0 & \sqrt{d^{row*}_{i+1}} \end{bmatrix} \cdot \begin{bmatrix} \bar{r}^{\circ}_{i,i} & \bar{r}^{\circ}_{i,i+1} \\ \bar{r}^{\circ}_{i+1,i} & \bar{r}^{\circ}_{i+1,i+1} \end{bmatrix} \\
&\qquad \cdot \begin{bmatrix} -\tan\phi\sqrt{\frac{d^{col}_{i}}{d^{col}_{i+1}}} & 1 \\ 1 & \tan\phi\sqrt{\frac{d^{col}_{i+1}}{d^{col}_{i}}} \end{bmatrix} \cdot \begin{bmatrix} \sqrt{d^{col}_{i+1}}\cos\phi & 0 \\ 0 & \sqrt{d^{col}_{i}}\cos\phi \end{bmatrix} \\
&= \begin{bmatrix} \sqrt{d^{row*}_{i}} & 0 \\ 0 & \sqrt{d^{row*}_{i+1}} \end{bmatrix} \cdot \begin{bmatrix} \bar{r}^{*}_{i,i} & \bar{r}^{*}_{i,i+1} \\ 0 & \bar{r}^{*}_{i+1,i+1} \end{bmatrix} \cdot \begin{bmatrix} \sqrt{d^{col*}_{i}} & 0 \\ 0 & \sqrt{d^{col*}_{i+1}} \end{bmatrix},
\end{aligned}$$

which then leads to column transformation formulas

$$\begin{bmatrix} \bar{r}^{*}_{i,i} & \bar{r}^{*}_{i,i+1} \\ 0 & \bar{r}^{*}_{i+1,i+1} \end{bmatrix} = \begin{bmatrix} \bar{r}^{\circ}_{i,i} & \bar{r}^{\circ}_{i,i+1} \\ \bar{r}^{\circ}_{i+1,i} & \bar{r}^{\circ}_{i+1,i+1} \end{bmatrix} \cdot \begin{bmatrix} -\frac{\bar{r}^{\circ}_{i+1,i+1}}{\bar{r}^{\circ}_{i+1,i}} & 1 \\ 1 & \frac{\bar{r}^{\circ}_{i+1,i+1} d^{col}_{i+1}}{\bar{r}^{\circ}_{i+1,i} d^{col}_{i}} \end{bmatrix}$$

with scale factor updating

$$\begin{aligned}
d^{col*}_{i} &= d^{col}_{i+1}\cos^2\phi, \\
d^{col*}_{i+1} &= d^{col}_{i}\cos^2\phi,
\end{aligned}$$

$$\cos^2\phi = \frac{1}{1 + \frac{\bar{r}^{\circ 2}_{i+1,i+1} d^{col}_{i+1}}{\bar{r}^{\circ 2}_{i+1,i} d^{col}_{i}}}.$$

5.2. Square root-free SVD updating. The above approximate SVD schemes straightforwardly combine with the square root-free QR updating procedure into a square root-free SVD updating procedure. At a certain time step, the data matrix is reduced to R, which is stored in factorized form

$$R = D^{\frac{1}{2}}_{row} \cdot \bar{R} \cdot D^{\frac{1}{2}}_{col}.$$

Furthermore, the same column scaling is applied to the V-matrix

$$V = \bar{V} \cdot D^{\frac{1}{2}}_{col},$$

where $\bar{V}$ is stored instead of V. The reason for this is twofold. First, the column rotations to be applied to the V-matrix are computed as modified Givens rotations. Explicitly applying these transformations to an unfactorized V would then necessarily require square roots. Second, a new row vector a^T to be updated immediately gets

TABLE 2

	SVD updating	Re-orthog.	Total
Internal cell	10×	12×	22×
V-matrix	10+	8+	18+
Internal cell	14×		14×
R-matrix	14+		14+
Diagonal cell	35×		35×
R-matrix	17+		17+
	9÷		9÷

the correct column scaling from the matrix-vector product $a^T \cdot \bar{V}$, so that the QR updating can then be carried out as if there were no column scaling at all. The updating can indeed be described as follows:

$$\begin{aligned}\begin{bmatrix} A \\ a^T \end{bmatrix} &= \begin{bmatrix} U & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} D_{row}^{\frac{1}{2}} \cdot \bar{R} \cdot D_{col}^{\frac{1}{2}} \\ a^T \cdot (\bar{V} \cdot D_{col}^{\frac{1}{2}}) \end{bmatrix} \cdot V^T \\ &= \begin{bmatrix} U & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} D_{row}^{\frac{1}{2}} & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \bar{R} \\ a^T \cdot (\bar{V}) \end{bmatrix} \cdot D_{col}^{\frac{1}{2}} \cdot V^T \\ &= \text{etc.}\end{aligned}$$

The factor in the midst of this expression can then be reduced to a triangular factor by QR updating, making use of modified Givens rotations. The further reduction of the resulting triangular factor can be carried out next, as detailed in the previous section.

From the above explanation, it follows that on a systolic array, a square root-free updating algorithm imposes hardly any changes. The diagonal matrices D_{row} and D_{col} are obviously stored in the processor elements on the main diagonal, and $\bar{R}$ and $\bar{V}$ are stored instead of R and V. The matrix-vector product $\bar{a}_\star^T = a^T \cdot \bar{V}$ is computed in the square part, and the $\bar{R}$-factor is updated with $\bar{a}_\star^T$ next, much like the R-factor was updated with $a_\star^T = a^T \cdot V$ in the original algorithm. All other operations are carried out much the same way, albeit that modified rotations are used throughout. When re-orthogonalizations are included, it is necessary to propagate the scale factors to the square array, along with the column transformation, such that the norms and inner products can be computed consistently. The rest is straightforward.

Finally, an operation count for a square root-free implementation is exhibited in Table 2. Note that the operation count for the diagonal processors depends heavily on the specific implementation. We refer to the literature for various (more efficient) implementations [1], [7]. The operation count for V-processors remains unchanged (as compared to Table 1), while the computational load for the diagonal processors is reduced to roughly the same level, apart from the divisions. The internal R-processors are seen to be underloaded this time, due to the reduction in the number of multiplications.

6. Conclusion. An approximate SVD updating procedure was mapped onto a systolic array with $O(n^2)$ parallelism for $O(n^2)$ complexity. By combining modified Givens rotations with approximate schemes for the computation of rotation angles in the SVD steps, all square roots can be avoided. In this way, a main computational bottleneck for the array implementation can be overcome.

REFERENCES

[1] J. L. Barlow and I. C. F. Ipsen, *Scaled Givens rotations for the solution of linear least squares problems on systolic arrays*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 716–733.

[2] R. P. Brent and F. T. Luk, *The solution of singular value and symmetric eigenvalue problems on multiprocessor arrays*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 69–84.

[3] J. P. Charlier, V. Vanbegin, and P. Van Dooren, *On efficient implementations of Kogbetliantz's algorithm for computing the singular value decomposition*, Numer. Math., 52 (1988), pp. 279–300.

[4] W. M. Gentleman, *Least squares computations by Givens transformations without square roots*, J. Inst. Math. Appls., 12 (1973), pp. 329–336.

[5] W. M. Gentleman and H. T. Kung, *Matrix triangularization by systolic arrays*, Real-Time Signal Processing IV, Proc. SPIE, Vol. 298, 1981, pp. 19–26.

[6] P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunders, *Methods for modifying matrix factorizations*, Math. Comp., 28 (1974), pp. 505–535.

[7] S. Hammarling, *A note on modifications to the Givens plane rotations*, J. Inst. Math. Appls., 13 (1974), pp. 215–218.

[8] M. T. Heath, A. J. Laub, C. C. Paige, R. C. Ward, *Computing the singular value decomposition of a product of two matrices*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1147–1159.

[9] E. Kogbetliantz, *Solution of linear equations by diagonalization of coefficient matrices*, Quart. Appl. Math., 13 (1955), pp. 123–132.

[10] F. T. Luk, *A triangular processor array for computing singular values*, Linear Algebra Appl., 77 (1986), pp. 259–273.

[11] F. T. Luk and H. Park, *On the equivalence and convergence of parallel Jacobi* SVD *algorithms*, in Proc. SPIE, Vol. 826, Advanced Algorithms and Architectures for Signal Processing II, F. T. Luk, ed., 1987 pp. 152–159.

[12] M. Moonen, P. Van Dooren, and J. Vandewalle, *An* SVD *updating algorithm for subspace tracking*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1015–1038.

[13] J. M. Speiser, *Signal processing computational needs*, in Proc. SPIE, Vol. 696, Advanced Algorithms and Architectures for Signal Processing, J. M. Speiser, ed., 1986, pp. 2–6.

[14] G. W. Stewart, *A Jacobi-like algorithm for computing the Schur decomposition of a nonhermitian matrix*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 853–863.

[15] S. Van Huffel and J. Vandewalle, *Algebraic relations between classical regression and total least squares estimation*, Linear Algebra Appl., 93 (1987), pp. 149–162.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 372-389, April 1993

004

AN ALGORITHM FOR THE BANDED SYMMETRIC GENERALIZED MATRIX EIGENVALUE PROBLEM*

LINDA KAUFMAN†

Abstract. This paper derives an algorithm for finding the eigenvalues of the symmetric banded generalized eigenvalue problem $A\mathbf{x} = \lambda B\mathbf{x}$ where A and B are $n \times n$ symmetric positive definite matrices of bandwidth $2k + 1$. Traditionally, for the nonbanded symmetric problem the Martin–Wilkinson algorithm has been used. This algorithm has been adapted to banded problems by Crawford and the current author has shown how to rearrange Crawford's algorithm to take advantage of parallel machines. Wang and Zhao have recently proposed another algorithm for the nonbanded symmetric generalized problem which seems to be able to compute small eigenvalues more accurately than the Martin–Wilkinson algorithm on problems with graded eigenvalues and ill-conditioned B. Their algorithm, like the Martin–Wilkinson algorithm, has a direct reduction phase requiring $O(n^3)$ operations followed by an iterative phase requiring $O(n^2)$ operations. In this paper the Wang–Zhao algorithm is reworked for banded matrices so that it requires $O(nk)$ space and $O(n^2k)$ operations, a reduction of $O(n/k)$ operations over the general algorithm. On a vector machine the new algorithm requires $O(nk^2)$ vector operations with vector lengths as long as n/k elements.

Key words. eigenvalues, banded, parallel computation

AMS(MOS) subject classification. 65F15

1. Introduction. In this paper we consider the generalized eigenvalue problem

$$A\mathbf{x} = \lambda B\mathbf{x}, \tag{1.1}$$

where A and B are both symmetric and positive definite $n \times n$ matrices and both are banded with $2k + 1$ nonzero diagonals.

There are two traditional approaches to solving (1.1) for general symmetric matrices (not banded). The first approach, suggested by Martin and Wilkinson [8], involves first finding the Cholesky factorization of B given by

$$B = L_B L_B^T, \tag{1.2}$$

where L_B is lower triangular, forming $C = L_B^{-1} A L_B^{-T}$, reducing it to a tridiagonal matrix C_T, and then finding the eigenvalues of C_T. Forming C_T is a direct method involving no iteration and requiring $O(n^3)$ operations, while finding its eigenvalues only is an iterative algorithm requiring $O(n^2)$ operations.

The other approach, the QZ algorithm of Moler and Stewart [9], disregards the symmetry of the problem, and after a direct phase requiring $O(n^3)$ operations, enters an iterative phase requiring $O(n^3)$ operations to simultaneously reduce A and B to triangular-quasi-triangular form. Because the Martin–Wilkinson algorithm explicitly uses the inverse of B to form C, their algorithm is much more sensitive to the conditioning of B than the QZ algorithm.

For banded matrices, Crawford [3] has suggested a method for modifying the Martin and Wilkinson algorithm so that the direct phase of the algorithm requires only $O(kn^2)$ operations and $O(nk)$ space. For tridiagonal matrices this means that the total operation count is $O(n^2)$ to find eigenvalues (i.e., no eigenvectors), which is much less than the $O(n^3)$ operations required by the QZ algorithm, which cannot take much advantage of the band structure in the matrices. Also, Kaufman has shown how to rearrange the planar

* Received by the editors March 4, 1991; accepted for publication (in revised form) September 4, 1991.

† Bell Laboratories, Murray Hill, New Jersey 07974 (lck@research.att.com).

rotations in Crawford's algorithm so that only $O(nk^2)$ vector operations are needed for a banded matrix. Therefore, for narrow bandwidth problems, the method used for determining the eigenvalues of the matrix C_T using, say IMTQL1 of EISPACK [12], completely dominates the operation count.

Recently, Wang and Zhao have proposed an algorithm for solving (1.1) for general symmetric positive definite matrices which requires about the same computational resources as the Martin–Wilkinson algorithm, i.e., much less than QZ. With graded problems with ill-conditioned B matrices, Kaufman has found that the smallest eigenvalues determined by the Wang–Zhao algorithm were much more accurate than those determined by the Martin–Wilkinson algorithm, but the largest eigenvalues determined by the Martin–Wilkinson algorithm were usually more accurate than those obtained by the Wang–Zhao algorithm. Thus the two algorithms complemented each other.

In this paper, we will consider applying the Wang–Zhao algorithm to banded matrices where their algorithm will have the greatest impact because for such matrices, using QZ can be rather expensive. In § 2 the Wang–Zhao algorithm is reviewed.

In § 3 the Wang–Zhao algorithm is modified for tridiagonal matrices. The new algorithm requires $O(n^2)$ operations. The modification is similar to Crawford's modification of the Martin–Wilkinson algorithm. The steps are arranged to take advantage of vector and parallel machines in a fashion similar to Kaufman's modification [6] of the Crawford algorithm.

In § 4 the ideas of § 3 are extended to a general banded matrix of bandwidth $2n + 1$. An algorithm is derived which requires $O(n^2k)$ operations and $O(nk)$ space. Thus we arrive at an algorithm that is rather insensitive to the ill-conditioning of B and that for narrow bandwidths requires about the same time to solve the standard eigenvalue problem with a symmetric tridiagonal matrix.

2. The algorithm of Wang and Zhao. Wang and Zhao [11] have recently proposed an algorithm for solving the eigenvalue problem

$$A\mathbf{x} = \lambda B\mathbf{x}, \tag{2.1}$$

where both A and B are symmetric and positive definite $n \times n$ matrices. Their algorithm is based on a theorem of Stewart [13] about the singular value decomposition (SVD) of portions of an orthogonal matrix. Wang and Zhao suggest that one consider the Cholesky factorizations of A and B given by

$$A = L_A L_A^T \tag{2.2}$$

and

$$B = L_B L_B^T, \tag{2.3}$$

and the QR decomposition

$$\begin{pmatrix} L_A^T \\ L_B^T \end{pmatrix} = QR, \tag{2.4}$$

where Q is a $2n \times n$ matrix with orthonormal columns and R is upper triangular and nonsingular, because of the positive definiteness of A and B. Now they suggest partitioning the Q of (2.4) into

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}, \tag{2.5}$$

where Q_1 and Q_2 each have n rows. Now from (2.4) and (2.5),

$$L_A^T = Q_1 R \tag{2.6}$$

and

$$L_B^T = Q_2 R. \tag{2.7}$$

If λ and $\mathbf{x}$ satisfy (2.1), then from (2.2) and (2.3)

$$L_A L_A^T \mathbf{x} = \lambda L_B L_B^T \mathbf{x},$$

which from (2.6) and (2.7) implies

$$R^T Q_1^T Q_1 R\mathbf{x} = \lambda R^T Q_2^T Q_2 R\mathbf{x}. \tag{2.8}$$

Now they apply Stewart's theorem [13], which says that if one obtains the SVDs

$$Q_1 = U_1 C V_1^T, \tag{2.9}$$

$$Q_2 = U_2 S V_2^T, \tag{2.10}$$

where U_1, U_2, V_1, and V_2 are all orthogonal matrices and C and S are diagonal matrices diag (c_i) and diag (s_i), respectively, then because of (2.5) and the orthogonality of the columns of Q in (2.4),

$$V_1 = V_2$$

and

$$c_i^2 + s_i^2 = 1. \tag{2.11}$$

If one lets $V = V_1$, then from (2.8), (2.9), and (2.10),

$$R^T V C^2 V^T R\mathbf{x} = \lambda R^T V S^2 V^T R\mathbf{x}. \tag{2.12}$$

Setting $\mathbf{y} = V^T R\mathbf{x}$ and multiplying both sides of (2.12) by $V^T R^{-T}$ yields

$$C^2 \mathbf{y} = \lambda S^2 \mathbf{y},$$

which, because C and S are diagonal, implies

$$\lambda_i = \frac{c_i^2}{s_i^2}. \tag{2.13}$$

From (2.11),

$$\lambda_i = \frac{c_i^2}{1 - c_i^2}. \tag{2.14}$$

From (2.2), (2.3), (2.4), (2.9), and (2.14), Wang and Zhao devised the following algorithm.

ALGORITHM WZ.

(1) Find the Cholesky factorizations

$$L_A L_A^T = \mathrm{A}, \qquad L_B L_B^T = B.$$

(2) Let

$$F = \begin{pmatrix} L_A^T \\ L_B^T \end{pmatrix}$$

and find orthogonal transformations $P_1, P_2, \ldots, P_n$ such that

$$P_n \cdots P_2 P_1 F = \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where R is upper triangular and P_i is designed to annihilate the elements below the diagonal of the ith column of $F^{(i)} = P_{i-1} \cdots P_2 P_1 F$.

(3) Determine

$$M = P_n \cdots P_2 P_1 \begin{pmatrix} I_n \\ 0 \end{pmatrix}.$$

Let $T^T = (I_n | 0) M$.

(4) Reduce T to bidiagonal form Q_B.

(5) Apply the iterative part of the SVD algorithm to Q_B to obtain singular values, $c_1, c_2, \cdots c_n$. The eigenvalues of (2.1) are then given by (2.14)

In [4], Golub and Pereyra encounter a problem similar to that in step (2) of Algorithm WZ and show that one can determine R and take advantage of the triangular matrices that make up F. They suggest that P_i should be a Householder transformation in planes i and $n + 1$ through $n + i$.

Steps 1–4 of the Wang–Zhao Algorithms are all $O(n^3)$ algorithms which can be easily vectorized. Step (5) is $O(n^2)$ but is usually implemented in a way which is not vectorizable (see [12]).

For singular values of (2.9) that are very close to 1, there could be a cancellation of significant figures when forming the eigenvalues using (2.14). In this case Wang and Zhao suggest using (2.10) and (2.13) rather than (2.14).

We note that Algorithm WZ requires that A be positive definite while the Martin–Wilkinson algorithm does not have this restriction. In an early version of their paper [10], Wang and Zhao suggest that this restriction can be overcome by replacing A by $A + \beta B$ and B by βB in step (1), and multiplying the right-hand side of (2.14) by β, but they did not give a completely satisfactory algorithm for determining β.

3. Modifying the Wang–Zhao algorithm to tridiagonal matrices.

In this section we show how to modify Algorithm WZ to take advantage of the structure of A and B when they are tridiagonal.

When Algorithm WZ is applied to tridiagonal matrices, the matrices L_A and L_B each have two diagonals and step (1) of Algorithm WZ requires only $O(n)$ operations. We are not as fortunate with the other steps of Algorithm WZ.

3.1. Decreasing the fill-in in F.

By studying more closely the tridiagonal problem, one can derive an algorithm for that case so that step (2) of Algorithm WZ requires $O(n)$ operations and step (3) requires $O(n^2)$ operations.

If one applied Algorithm WZ to a tridiagonal problem using a single Householder transformation to annihilate elements below the subdiagonal of F in step (2), then the last n rows of $F^{(3)}$ would look like

$$(3.1)\qquad \begin{pmatrix} 0 & 0 & + & & & \\ & 0 & x & & & \\ & & x & x & & \\ & & & x & \ddots & \\ & & & & \ddots & x \\ & & & & & x \end{pmatrix},$$

where "+" indicates a new nonzero element.

If one constructed $F^{(4)}$ using a single Householder transformation, as one would in Algorithm WZ for general matrices, then the last n rows of $F^{(4)}$ would look like

$$(3.2)\qquad \begin{pmatrix} 0 & 0 & 0 & + & & \\ & 0 & 0 & + & & \\ & & 0 & x & & \\ & & & x & \ddots & \\ & & & & \ddots & x \\ & & & & & x \end{pmatrix}$$

with fill-in in both the $(n + 1, 4)$ and $(n + 2, 4)$ positions. In general, on a tridiagonal matrix, using a single Householder to eliminate elements in the ith column requires affecting planes i and $n + 1$ through $n + i$ so that $O(n^2)$ operations would be required for step (2) and $O(n^3)$ operations would be required for step (3).

It is more economical, however, if P_i is not a single Householder transformation but is composed of transformations which only affect a few rows at a time. Let us denote by G_i a transformation in planes i and $i + 1$. G_i could be a Givens rotation or a Householder transformation. Let us denote by H_i a transformation in planes i, $n + i$, and $n + i - 1$ designed to eliminate elements in planes $n + i$ and $n + i - 1$. The product $P_3 = H_3G_{n+1}$, where G_{n+1} is designed to annihilate $f^{(3)}_{n+1,3}$ and H_3 is designed to annihilate the remaining elements in the third column below the diagonal, would produce the matrix $F^{(4)}$ whose last n rows would look like

$$(3.3)\qquad \begin{pmatrix} 0 & 0 & 0 & & & \\ & 0 & 0 & + & & \\ & & 0 & x & & \\ & & & x & \ddots & \\ & & & & \ddots & x \\ & & & & & x \end{pmatrix}$$

with fill-in only in the $(n + 2, 4)$ position. If one continued in this mode of using a planar rotation G_{n+i-2} to annihilate $f^{(i)}_{n+i-2,i}$, followed by H_i to annihilate the other elements below the subdiagonal to form $F^{(i+1)}$, then step (2) would require $O(n)$ operations for a tridiagonal matrix and step (3) would require $O(n^2)$ operations.

3.2. Decreasing the fill-in in M. Thus far we have considered decreasing the fill-in in F, but we have not approached the problem that because T is a filled-in triangular matrix in step (3), step (4) requires $O(n^3)$ operations and $O(n^2)$ space for banded matrices. To handle that problem, we devise an algorithm similar to that given by Crawford

[3] to modify the Martin–Wilkinson algorithm for the symmetric generalized eigenvalue problem in order to take advantage of band matrices. As in § 3.1, we will only consider tridiagonal matrices.

In step (3) of Algorithm WZ, let

$$M^{(i)} = P_{i-1} \cdots P_2 P_1 \begin{pmatrix} I_n \\ 0 \end{pmatrix}. \tag{3.4}$$

Because our original matrices are tridiagonal, $M^{(4)} = H_3 G_{n+1} M^{(3)}$, where G_{n+1} is a two-planar transformation designed to annihilate $f_{n+1,3}^{(3)}$. Except for the "+" element, $M^{(4)}$ will look like

$$\begin{pmatrix} x & + & & & & \\ x & x & & & & \\ e & f & f & & & \\ & & & 1 & & \\ & & & & \cdot & \\ & & & & & 1 \\ x & x & & & & \\ d & x & x & & & \\ e & f & x & & & \\ & & & 0 & & \\ & & & & \cdot & \\ & & & & & 0 \end{pmatrix}. \tag{3.5}$$

Future transformations will fill in the lower triangular part of the first n rows of the M matrices. The source of the problem is the $(n+1, 1)$ element in $G_{n+1}M^{(3)}$, marked with a d in (3.5). If that element were 0, the first three rows of (3.5) would be bidiagonal.

However, we can make the $(n+1, 1)$ element of $G_{n+1}M^{(3)}$ zero by using a column transformation in the (1, 2) plane, which will add the "+" element pictured in (3.5). This fill-in can be eliminated by a two-plane transformation in the (1, 2) plane. At this point applying H_3 will keep the first three rows of $M^{(4)}$ bidiagonal. The e elements in (3.5) will never appear.

In fact, for tridiagonal matrices before H_i is applied, one can do column transformations involving the first $i-1$ columns and row transformations in the first $i-1$ rows to get rid of unwanted fill without affecting later computations. By chasing unwanted elements up the matrix using two-planar rotations it is possible to keep the first n rows of all the M matrices bidiagonal, so step (4) of Algorithm WZ will be unnecessary.

The matrix $M^{(i)}$ will have the structure

$$\begin{pmatrix} W & & & 0 \\ 0 & & & I_{n-i+1} \\ U & & & \\ 0 & x & x & 0 \\ 0 & + & x & \\ & & 0 & \end{pmatrix}, \tag{3.6}$$

where W is an $(i-1) \times (i-1)$ *bidiagonal* matrix and U is an $(i-3) \times (i-1)$ matrix.

Applying G_{n+i-2} to $M^{(i)}$ will not change its structure. However, applying H_i immediately will cause an unwanted element in the $(i, i-2)$ position. To prevent this, before one applies H_i, a *column* transformation G_{i-2} is used to annihilate the "+" element in (3.6). Applying the column transformation to the current M matrix results in unwanted fill in the $(i-2, i-1)$ position which can be chased up the diagonal. The chasing operation is depicted in (3.7) for a small example. Only one numbered entry is nonzero at any one time and the number gives the order in which they appear and are eliminated:

$$(3.7) \qquad \begin{pmatrix} x & 7 & & & \\ x & x & 5 & & \\ 6 & x & x & 3 & \\ & 4 & x & x & 1 \\ & & 2 & x & x \end{pmatrix}.$$

Although the column transformations used in the chase in (3.7) will change the U matrix in (3.6), the rows of the U matrix never affect the rest of the algorithm to determine T. Thus there is no reason to store the U portion of the M matrices or to apply the transformations to it. Thus all the transformations on the M matrices are applied to vectors of length 3 at most. Since eliminating the ith column involves about $2i$ transformations for the chase, the total algorithm to compute T involves $O(n^2)$ transformations and requires $O(n^2)$ operations. We will call this chasing scheme algorithm TRI.

3.3. A parallel algorithm. Each of the $n-1$ chases depicted in (3.7) is a serial operation. However, the G and H operations designed to annihilate the elements of F and the column transformation designed to annihilate the "+" in (3.6) is not dependent on the *completion* of the previous chases. All that is required is that the previous chases have progressed far enough so that they do not interfere with the beginning of the next chase. Thus one can have several simultaneous chases operating simultaneously, i.e., in parallel, as in Kaufman's modification [6] of Crawford's scheme, as long as they do not interfere with each other. In (3.8) assume that the x's and a's are nonzero:

$$(3.8) \qquad \begin{pmatrix} x & d & & & & & & & \\ x & x & b & & & & & & \\ c & x & x & d & & & & & \\ & a & x & x & b & & & & \\ & & c & x & x & d & & & \\ & & & a & x & x & b & & \\ & & & & c & x & x & d & \\ & & & & & a & x & x & \\ & & & & & & c & x & x \end{pmatrix}.$$

Doing row Givens transformations to eliminate all the a's will produce the b's. Since these row transformations access distinctly different rows, they may occur simultaneously. Similarly, the b's can be turned into c's with similar column transformations. A new c can now be introduced at the bottom because of manipulations of the F matrix. Now simultaneous row transformations can change the c's into d's. The top d can be chased out the top and the rest turned into a's, beginning the process again.

For general n one will eventually be moving to $(n-1)/2$ elements up the superdiagonal of the M matrix. The basic step involves simultaneously determining many Givens transformations and then simultaneously applying these transformations to a few vectors. The operation is a lockstep one that is appropriate to a single instruction, multiple data (SIMD) machine. Since the data can be arranged so that operands are within a fixed stride of each other, one can take advantage of the architecture of a vector machine. Rather than characterizing steps (1)–(3) of Algorithm WZ for tridiagonal matrices as requiring $O(n^2)$ operations, one can describe the methods as requiring $O(n)$ vector operations as long as one can convince the compiler on one's parallel machine or vector machine to compute Givens transformation in parallel.

In Algorithm TRIP, which summarizes the whole parallel algorithm including the chase, the superscripts are removed from F and M for simplicity.

ALGORITHM TRIP.

```
Eliminate $f_{n+1,1}$ using a row transformation in planes 1 and $n+1$.
Determine $H_2$ to eliminate $f_{n+1,2}$ and $f_{n+2,2}$ introducing $f_{n+1,3}$ as in (3.1).
For $i = 3, \ldots, n$
    (a) Eliminate $f_{n+i-2,i}$ by a row transformation $G_{n+i-2}$ as in (3.3).
    (b) Eliminate $m_{n+i-1,i-2}$ by a column transformation $G_{i-2}$ introducing
        $m_{i-2,i-1}$ as in (3.7).
    (c) For $j = i - 1$ step $-2$ until 1
            Eliminate $m_{j-1,j}$ using a row transformation $G_{j-1}$ introducing $m_{j,j-2}$ as
            in (3.8).
    (d) For $j = i - 1$ step $-2$ until 3
            Eliminate $m_{j,j-2}$ using a column transformation $G_{j-2}$ introducing $m_{j-2,j-1}$
            as in (3.8).
    (e) Eliminate $f_{n+i-1,i}$ and $f_{n+i,i}$ using $H_i$ as in (3.3).
```

Note that in Algorithm TRIP one does not actually have to save M and F as full matrices. Four vectors of length n are sufficient to store F. Since there is no need to store the U portion of the M matrices in (3.6) and the W matrix is tridiagonal, three vectors of length n suffice to store M.

The first two columns of Table 3.1 summarize the operation count for general matrices and Algorithm TRI. To a certain extent, because the operation count is bounded by the SVD algorithm in step (5), there seems little reason to work further on the first three steps for tridiagonal matrices.

In Table 3.2, we compare computation times for $n = 100$ and $n = 200$ on the SGI 4D/240, a fast scalar machine, and on the Cray XMP for various alternatives for solving $A\mathbf{x} = \lambda B\mathbf{x}$. For the SGI machine the times are accurate to within 20 percent and for the CRAY, to within about 10 percent. We see that reducing the fill-in in T was absolutely essential to remain competitive on the SGI. The reduction of a matrix to bidiagonal form is a costly operation that, thankfully, Algorithm TRI avoids. Since the iterative part of the SVD algorithm is slightly more expensive than IMTQL1 of [12], Crawford's approach is slightly cheaper. On a vector machine like the Cray, the case for taking advantage of band structure is not as strong from the point of view of execution time, because the general algorithms can be easily vectorized. For Table 3.2 compiler directives were used to insure vectorization on the Cray XMP for Algorithm TRIP.

TABLE 3.1
Operation count for Algorithms WZ *and* TRI.

Step	Algorithm WZ for general matrices	Algorithm TRI	Algorithm TRI for banded matrices with bandwidth $2k + 1$
1	$O(n^3)$	$O(n)$	$O(nk)$
2	$O(n^3)$	$O(n)$	$O(nk^2)$
3	$O(n^3)$	$O(n^2)$	$O(n^2k)$
4	$O(n^3)$	0	$O(n^2k)$
5	$O(n^2)$	$O(n^2)$	$O(n^2)$

TABLE 3.2
Execution times on a tridiagonal matrix.

Algorithm	SGI $n = 100$	SGI $n = 200$	Cray XMP $n = 100$	Cray XMP $n = 200$
RGG [12]	22	84	.39	2.1
RSG [12]	13	57	.12	.55
Crawford to tridiagonal	.12	.50	.019	.078
IMTQL1 [12]	.17	.62	.035	.133
Total	.29	1.12	.054	.211
Alg. TRI (SGI), TRIP (Cray)	.15	.63	.0057	.0176
SVD, step 5	.23	.83	.069	.268
Total	.37	1.46	.075	.286

For $n = 200$, on the Cray XMP Algorithm TRIP was about 7 times faster than Algorithm TRI.

3.4. Numerical experiments. Our numerical experiments were not conclusive. Several examples were run on the SGI machine using Crawford's scheme, RSG from Eispack, the Martin–Wilkinson code, and RGG, the QZ code, in double precision. The results from RGG in double precision should be considered accurate. As theory suggests, the errors in the Martin–Wilkinson-like algorithms tended to be small relative to the largest eigenvalue of the problem. Thus, when the eigenvalues were distributed in a range that was about the same or larger than the inverse of the machine precision, the smaller eigenvalues computed by the Martin–Wilkinson-like scheme was often very wrong, but the larger eigenvalues were fine. With our code, eigenvalues that were near the inverse of machine precision were often far off, but the small eigenvalues were accurate. Thus any comparison has to be dependent on the user's goals and priorities.

Table 3.3 considers our first example in which the B matrix was constructed by choosing a diagonal matrix D as $D_{ii} = 10^{(i-1)}$, applying a similarity Householder matrix to get a dense matrix and then reducing that matrix to tridiagonal form. This way we could insure that we had an ill-conditioned tridiagonal matrix. For the matrix A, the diagonal elements were set to 1 and the off-diagonal elements to .01. In Table 3.3 and

TABLE 3.3
Numerical results for example 1.

	Relative error in			Angles		
Eigenvalues	TRI	Crawford	RSG	TRI	Crawford	RSG
1.0098E−09	8.97E−05	5.69E−06	8.42E−01	9.06E−14	5.75E−15	7.81E−10
1.0014E−08	3.90E−04	2.38E−06	1.05E−01	3.90E−12	2.39E−14	1.05E−09
9.9676E−08	1.64E−04	3.93E−06	7.60E−04	1.64E−11	3.92E−13	7.57E−11
9.9600E−07	1.24E−04	2.87E−06	3.41E−05	1.24E−10	2.86E−12	3.40E−11
9.9600E−06	9.95E−05	2.34E−05	1.00E−05	9.91E−10	2.33E−10	9.98E−11
9.9668E−05	1.75E−04	3.16E−04	8.55E−05	1.74E−08	3.14E−08	8.53E−09
9.9693E−04	9.57E−04	3.14E−03	8.71E−04	9.54E−07	3.13E−06	8.68E−07
8.5414E−03	9.91E−03	3.55E−02	9.80E−03	8.43E−05	3.02E−04	8.36E−05
6.0187E−02	8.66E−02	2.15E−01	8.65E−02	5.04E−03	1.26E−02	5.17E−03
1.8479E−01	1.12E−02	2.31E−02	1.12E−02	1.87E−03	3.87E−03	1.98E−03

TABLE 3.4
Numerical results for example 2 *with* $n = 7$.

	Relative errors in			Angles		
Eigenvalues	TRI	Crawford	RSG	TRI	Crawford	RSG
8.6448E−05	2.26E−04	4.31E−02	2.23E−02	1.96E−08	3.73E−06	1.93E−06
5.0791E−03	4.09E−06	2.38E−04	2.29E−03	2.07E−08	1.21E−06	1.16E−05
1.0175E−01	7.49E−08	2.42E−06	2.59E−04	7.22E−09	2.33E−07	2.50E−05
1.0000E+00	1.32E−06	1.68E−06	2.18E−04	6.61E−07	8.39E−07	1.09E−04
9.8284E+00	5.73E−06	6.70E−07	2.03E−05	1.73E−06	2.03E−07	6.13E−06
1.9688E+02	1.43E−07	7.63E−07	1.38E−06	1.02E−08	5.43E−08	9.83E−08
1.1561E+04	4.76E−03	4.68E−05	2.06E−04	4.44E−05	4.35E−07	1.91E−06

for the other numerical experiments in this paper we give not only the relative errors in the eigenvalues but also the associated angles as described in Stewart and Sun [14].

Perhaps the most surprising fact about Table 3.3 is that the results for RSG and the Crawford algorithm differed more from each other than they differed from TRI. Since the Crawford algorithm is supposedly a banded version of RSG, this was not expected. Closer examination of the code suggested that the difference was caused by the fact that in RSG the Cholesky factorization was begun at the top, and that in the Crawford algorithm the Cholesky factorization started at the bottom of the B matrix, as well as the fact that the Crawford algorithm was a chasing scheme that used many Given's rotations, as opposed to a few Householder transformations. The results for this table suggest that for the smaller eigenvalues, TRI somewhat matched the Crawford scheme, while the Crawford scheme was better than RSG; but for the larger eigenvalues, TRI matched RSG, while RSG was superior to Crawford. In other words, TRI did very well on both ends of the spectrum, while the other two algorithms favored one end or the other.

When there was a large eigenvalue near the inverse of the machine precision the results were not as favorable as our second example indicates. In Tables 3.4 and 3.5 we began by creating a matrix C such that $c_{i,j} = \sin((i-j)^*\pi/2)/(i-j)$ for $i \geqq j$ and then setting $c_{j,i} = c_{i,j}$ for $i > j$. The matrix C was then reduced to tridiagonal form and this became our matrix B. The matrix A was then set to $\pi I - B$. We realize that there is an easier way to solve this problem than setting up a generalized eigenvalue problem, but we wished to test our algorithm on an example that had large and small eigenvalues.

Tables 3.4 and 3.5 indicate our results when $n = 7$ and $n = 8$, respectively. Notice that for the smaller eigenvalues, Algorithm TRI does surprisingly well, but for the larger

TABLE 3.5
Numerical results for example 2 *with* $n = 8$.

	Relative errors in			Angles		
Eigenvalues	TRI	Crawford	RSG	TRI	Crawford	RSG
1.6303E−05	1.85E−04	5.70E+00	2.61E−01	3.02E−09	9.29E−05	4.25E−06
1.1395E−03	1.15E−05	6.90E−03	5.82E−02	1.32E−08	7.86E−06	6.63E−05
2.9387E−02	1.02E−06	2.82E−04	7.05E−03	2.95E−08	8.16E−06	2.04E−04
3.4739E−01	1.62E−07	7.81E−07	5.45E−04	4.59E−08	2.21E−07	1.54E−04
2.8786E+00	7.95E−08	5.76E−07	1.14E−04	3.81E−08	2.76E−07	5.48E−05
3.4029E+01	2.80E−05	3.42E−06	1.70E−05	4.72E−06	5.78E−07	2.87E−06
8.7764E+02	2.73E−05	8.20E−06	1.38E−05	9.20E−07	2.77E−07	4.64E−07
6.1911E+04	3.22E−02	1.24E−03	1.03E−03	1.32E−04	4.98E−06	4.14E−06

eigenvalues, Algorithm TRI loses. This should be expected considering (2.14). For large singular values, squaring them to obtain the eigenvalue values will amplify any error. Moreover, there will be a further cancellation error for a singular value close to 1. If one knew that this was the situation, one could also compute Q_2 and its SVD as in (2.10) and then use (2.13). Doing this computation would double the work.

When the algorithm of § 3.1, which does not involve chasing, was applied to the second example with $n = 8$, the five smallest eigenvalues were left unchanged, but the largest three were modified. The largest eigenvalue computed had an error of 1.82×10^{-2}, which shows how sensitive to roundoff error Algorithm WZ is for singular values near 1.

4. The general banded case. In this section we show how to modify Algorithm WZ to take advantage of the structures of A and B when they are banded with bandwidth $2k + 1$.

For general banded matrices of bandwidth $2k + 1$, it is sufficient to treat L_A and L_B as block bidiagonal matrices having at most $(\lfloor n/k \rfloor + 1)$ row blocks, each having a block size of k. For example, the matrix L_A^T with nine rows and $k = 3$ might be considered as

$$\begin{array}{|ccc|ccc|ccc|}
\hline
x & x & x & x & & & & & \\
 & x & x & x & x & & & & \\
 & & x & x & x & x & & & \\
\hline
 & & & x & x & x & x & & \\
 & & & & x & x & x & x & \\
 & & & & & x & x & x & x \\
\hline
 & & & & & & x & x & x \\
 & & & & & & & x & x \\
 & & & & & & & & x \\
\hline
\end{array}$$

One then applies Algorithm TRI in a block manner. This means that a transformation G will not be a planar transformation designed to zero one element, but be a transformation involving $2k$ rows designed to annihilate a $k \times k$ block. Similarly, an H transformation will work on $3k$ rows and be designed to annihilate a block that is $2k \times k$. These transformations may be generalized Givens transformations (see Bartels and Kaufman [1]), generalized Householder transformations (see Kaufman [7] or Bischof and Van Loan [2]), or just a sequence of single transformations lumped together. It is inconsequential that the resulting matrix R of (2.4) may not be upper triangular. However, it is consequential that the resulting T matrix from step (3) will not be bidiagonal, as in the tridiagonal case, but be block bidiagonal. By alternating block column transformations designed to eliminate the strictly upper triangular portions of the diagonal blocks of T with block row transformations designed to eliminate the strictly lower triangular portions of the off-diagonal blocks, T can be reduced to a lower triangular banded matrix. At this point Tsao's algorithm [15] can be used to reduce T^T to bidiagonal form.

4.1. Preserving band structure. However, the block bidiagonal approach hides the fact that one can preserve band structure. Specifically, the diagonal blocks of the top part of the M matrices can be kept lower triangular and the subdiagonal blocks can be kept upper triangular. Let us assume we have just completed a chasing operation leaving us with a lower triangular matrix with four nonzero diagonals. Before the next chase, the structure is ruined slightly as follows:

$$\tag{4.1} \begin{pmatrix} x &&&&&&&&&&& \\ x & x &&&&&&&&&& \\ x & x & x &&&&&&&&& \\ x & x & x & x &&&&&&&& \\ & x & x & x & x &&&&&&& \\ && x & x & x & x &&&&&& \\ &&& x & x & x & x & x & x & + & b & b \\ &&&& x & x & x & x & x & + & + & b \\ &&&&& x & x & x & x & + & + & + \\ &&&&&& x & x & x & x & + & + \\ &&&&&& x & x & x & x & x & + \\ &&&&&& x & x & x & x & x & x \end{pmatrix}.$$

The "b" elements can be eliminated using column transformations within the last filled-in block. These transformations should be applied to the bottom part of M also. Eliminating the "+" elements begins a chasing procedure which is depicted in (4.2). All the elements with the same number are eliminated in one group and the groups are processed in numerical order. Column transformations are used to eliminate "odd"-numbered elements and row transformations are used to eliminate even-numbered elements.

$$\tag{4.2} \begin{pmatrix} x & 6 & 6 & 4 &&&&&&&& \\ x & x & 6 & 4 & 4 &&&&&&& \\ x & x & x & 4 & 4 & 4 &&&&&& \\ x & x & x & x & 4 & 4 & 2 &&&&& \\ 5 & x & x & x & x & 4 & 2 & 2 &&&& \\ 5 & 5 & x & x & x & x & 2 & 2 & 2 &&& \\ 3 & 3 & 3 & x & x & x & x & 2 & 2 &&& \\ & 3 & 3 & 3 & x & x & x & x & 2 &&& \\ && 3 & 3 & 3 & x & x & x & x &&& \\ &&& 1 & 1 & 1 & x & x & x & x && \\ &&&& 1 & 1 & 1 & x & x & x & x & \\ &&&&& 1 & 1 & 1 & x & x & x & x \end{pmatrix}$$

Thus the band structure can be preserved during the elimination process.

Since we have $(n/k)^2$ block transformations for the chasing algorithm and each block transformation could involve $O(k^3)$ operations, the operation count for step (3) is now $O(n^2k)$. To store the top part of the M matrix, $(k + 1) \times n$ elements are needed.

Since the final result of applying the algorithm just described is a triangular matrix with $k + 1$ diagonals, step (4) of Algorithm WZ cannot be ignored as it was in the tridiagonal case. Tsao [15] has given an algorithm to reduce this type of matrix to bidiagonal form using $O(n^2k)$ operations.

The chasing algorithm described in (4.2) is more of a block chase, which is good for problems of large bandwidth. However, for problems with very narrow bandwidth it may be advantageous to abandon a block idea. In (4.1) one could eliminate the elements below the $k + 1$st subdiagonal and then use Givens transformations to chase the elements above the diagonal one at a time up the diagonal, as in Tsao's algorithm. The advantage of using Givens transformations is that one would always be concerned with two consecutive rows and columns, which would simplify the indexing.

The operation count for the banded case is summarized in the third column of Table 3.1. Our experience on the SGI machine with an implementation of the ideas in this section is given in Table 4.1. The data corroborates the theory that for all steps of the algorithm, doubling n increases the time by a factor of 4. There is so much overhead in the implementation of steps (1)–(3), that for small values of k, the time does not increase. Eventually however, it is evident that the times increase linearly with k.

4.2. A parallel algorithm. In the algorithm described in § 4.1 there are two instances of a serial chasing algorithm. The chase described in (4.2) can be changed into a parallel chasing algorithm, similar to that for the tridiagonal case, using at most $n/(2k)$ simultaneous block transformations. Thus the speedup would depend on how well a particular compiler could handle nested parallel operations. The second instance is the chasing algorithm in Tsao's algorithm for reducing a general triangular banded matrix to bidiagonal form.

4.2.1. A parallel version of Tsao's algorithm for two superdiagonals. We now show how to implement Tsao's algorithm [15] for this reduction, requiring $O(n^2k)$ operations, using $O(nk^2)$ vector operations. We begin with the case when there are two superdiagonals, an unfortunate special case.

Tsao's algorithm may be pictured as follows:

$$\begin{pmatrix}
x & x & x & h & & & & & & & & \\
i & x & x & x & 8 & & & & & & & \\
 & 9 & x & x & x & f & & & & & & \\
 & & g & x & x & x & 6 & & & & & \\
 & & & 7 & x & x & x & d & & & & \\
 & & & & e & x & x & x & 4 & & & \\
 & & & & & 5 & x & x & x & b & & \\
 & & & & & & c & x & x & \gamma & 2 & \\
 & & & & & & & 3 & x & x & \beta & \\
 & & & & & & & & a & x & x & \alpha \\
 & & & & & & & & & 1 & x & x \\
 & & & & & & & & & & & x
\end{pmatrix}. \tag{4.3}$$

TABLE 4.1
Execution times on the SGI *for the banded algorithm.*

n	k	RSG	Steps (1)–(3)	Step (4)	Step (5)	Total (1)–(5)
100	1	12	.65	0.	.25	.9
	2	12	.65	.13	.25	1.1
	4	12	.88	.27	.25	1.4
	8	12	1.6	.48	.25	2.3
	16	12	3.4	.80	.25	4.5
200	1	57	2.7	0.0	.83	3.5
	2	57	2.7	.55	.83	4.1
	4	57	3.5	1.1	.83	5.4
	8	57	5.5	2.0	.83	8.3
	16	57	11.1	3.5	.83	15.4

In (4.3) the original matrix consists of the x's, α, β, and γ. The elimination of α starts a chase up the diagonal in which unwanted elements are chased from position 1 to position 2 to position 3, etc. After the initial chase, β is eliminated filling up position a, which when eliminated produces an element in position b, etc., through the alphabet. After the second case, Tsao's algorithm returns to eliminate γ, which puts a nonzero in position 3 again and that element is chased through the numbered positions up. Notice that a chase hits every second element on the subdiagonal.

As in step (3) one does not have to complete a whole chase sequence before beginning the next in Tsao's algorithm. One can stop the first chase at position 5, eliminate β producing something at position a, and then simultaneously eliminate elements at positions 5 and a. Then one can do successively simultaneous eliminations at positions 6 and b, 7 and c, and 8 and d. If at this point one went back to eliminate γ producing a new element at position 3, one could simultaneously eliminate elements at positions 9, e, and 3 of (4.3). It should be apparent that eventually from the top of the matrix one would be doing simultaneous eliminations with every third subdiagonal element. However, because by the time the first elimination has reached the top subdiagonal, some of the bottom subdiagonals have been completely eliminated, the maximum number of simultaneous eliminations is less than $n/3$ and simple algebra shows that it is $n/4$. The following algorithm summarizes this paragraph.

ALGORITHM BIDIAGP FOR THREE DIAGONAL BANDED MATRICES.

```
Set $l \leftarrow n-2$
For $i=n-2$ step $-1$ until 1
    Eliminate $t_{i,i+2}$ with a row transformation $G_i$ introducing $t_{i+1,i}$.
    For $j=i$ step $-3$ until $l$
        Eliminate $t_{j+1,j}$ by a column transformation $G_j$ introducing $t_{j-2,j+1}$.
    For $j=i$ step $-3$ until max$(l,3)$
        Eliminate $t_{j-2,j+1}$ by a row transformation $G_{j-2}$ introducing $t_{j-1,j-2}$.
    For $j=i$ step $-3$ until max$(l,3)$
        Eliminate $t_{j-1,j-2}$ by a column transformation $G_{j-2}$ introducing $t_{j-4,j-1}$.
    For $j=i$ step $-3$ until max$(l,5)$
        Eliminate $t_{j-4,j-1}$ by a row transformation $G_{j-4}$ introducing $t_{j-3,j-4}$.
    $l \leftarrow$ max$(1,l-4)$
```

Table 4.2 compares Algorithm BidiagP with Tsao's algorithm on the Cray XMP using the CFT77 compiler with vectorization assured by using compiler directives. The Cray XMP was run in single processor mode. Certainly, doing simultaneous Givens transformations affected the computation time. As one would expect with Tsao's algorithm, the running time quadrupled when n was doubled. In the parallel algorithm, doubling n did not produce the same result. While BidiagP was aimed at rectangular matrices with three diagonals, Tsao's algorithm permits more diagonals, and thus one can expect some overhead because of the generality of the code.

TABLE 4.2
Comparison of BidiagP and Tsao's algorithm on the Cray XMP.

Algorithm	$n = 100$	$n = 200$
Tsao	.066	.265
BidiagP	.0043	.011

4.2.2. More than two superdiagonals. Algorithm BidiagP is made more complicated by the fact that in the chase in the Tsao algorithm every other element subdiagonal is hit. When there are k superdiagonals ($k > 2$) and one chooses to eliminate the uppermost diagonal at once, then every kth subdiagonal element is hit. This is illustrated in (4.4), with $k = 3$:

$$\begin{pmatrix}
x & x & x & x & 6 & & & & & & & & \\
7 & x & x & x & x & D & & & & & & & \\
 & E & x & x & x & x & d & & & & & & \\
 & & e & x & x & x & x & 4 & & & & & \\
 & & & 5 & x & x & x & x & B & & & & \\
 & & & & C & x & x & x & x & b & & & \\
 & & & & & c & x & x & x & \delta & 2 & & \\
 & & & & & & 3 & x & x & x & \gamma & & \\
 & & & & & & & A & x & x & x & \beta & \\
 & & & & & & & & a & x & x & x & \alpha \\
 & & & & & & & & & 1 & x & x & x \\
 & & & & & & & & & & & x & x \\
 & & & & & & & & & & & & x
\end{pmatrix}. \tag{4.4}$$

The x's, α, β, γ, and δ are elements of the original T matrix. Eliminating α produces a chase which successively creates nonzeros in positions 1, 2, 3, 4, . . . , etc. Eliminating β then produces a chase which successively creates and annihilates positions in positions $a, b, c, \ldots$, and eliminating γ produces a chase which successively creates and annihilates position in positions $A, B, C, \ldots$. Finally, eliminating δ puts something back in position 3 and the numbered positions are hit again.

As in the two-superdiagonal case, there is no need to complete the chase before beginning the next one. After α is eliminated, we need only chase the unknown to position 3. We can then eliminate β putting a nonzero in position a and then simultaneously annihilate the elements in positions 3 and a with column transformations putting elements in 4 and b, which can be annihilated with row transformations. At this point one can return to annihilate γ putting an unwanted element in position A. Now those in positions 5, c, and A can be annihilated by a column row combination and moved up to 7, e, and C in (4.4). Finally, eliminating δ creates a new nonzero at 3 and the nonzero at 7, e, C, and 3 can be eliminated simultaneously. Thus for $k = 3$, we enter the situation in which we are simultaneously removing every other element on the subdiagonal, a much more favorable situation then for two superdiagonals. In fact, the maximum vector length will be $n/3$ for $k = 3$, rather than $n/4$ for $k = 2$.

The above chase can be summarized as follows.

ALGORITHM REDUCEKP TO ELIMINATE ONE SUPERDIAGONAL.

```
Set $l \leftarrow n-k$
For $i=n-k$ step $-1$ until 1
    Eliminate $t_{i,i+k}$ with row transformation $G_i$ generating $t_{i+1,i}$.
    For $j=i$ step $-(k-1)$ until $l$
        Eliminate $t_{j+1,j}$ with a column transformation $G_j$ introducing $t_{j-k,j+1}$.
    For $j=i$ step $-(k-1)$ until max$(l,k+1)$
        Eliminate $t_{j-k,j+1}$ with a row transformation $G_{j-k}$ introducing $t_{j-k+1,j-k}$.
    Set $l \leftarrow$ max$(1,l-k)$
```

Algorithm ReducekP only peels off one superdiagonal. If the original matrix had k superdiagonals, one would call the ReducekP algorithm $k - 2$ times, each time reducing k, and then call BidiagP to finally get down to bidiagonal form. The reason one would work by diagonals, rather than by columns, is that vector lengths increase for smaller k until $k = 3$.

Table 4.3 compares Tsao's algorithm with ReducekP in single processor mode on the Cray XMP using the CFT77 compiler and compiler directives to assure vectorization. In Tsao's algorithm the inner loop is over the number of superdiagonals k, and thus for a given n, as k becomes somewhat large, the computation time does not appreciably increase. For ReducekP the inner loop is over the number of simultaneous Givens transformations which is inversely proportional to k. Thus as k increases, for a given n the lengths of the inner loops decrease and the total computation time would increase far faster than with Tsao's algorithm, since they are actually doing the same amount of arithmetic. Obviously there is a crossover point between the two algorithms. Also, comparing Tsao's algorithm with the information for the SVD on the Cray XMP given in Table 3.2, it seems that from a time point of view, one would never wish to use Tsao's algorithm without parallelizing it, as in ReducekP.

In this section we have shown how to decrease the first four steps in the banded algorithms of § 3 from $O(n^2k)$ operations to approximately (nk^2) vector operations where vector lengths are essentially $O(n/k)$. Thus for narrow banded systems the work involved on a vector machine for the first four steps of Algorithm WZ is essentially negligible. The work is dominated by one's implementation of the SVD algorithm which is independent of k.

4.3. Numerical experiments. The banded algorithm and RSG from Eispack [12] were applied to the following problem with an ill-conditioned B matrix: Set the matrix B to

$$b_{i,j} = \begin{cases} 0 & \text{if } |i-j| > k, \\ n \times 5^{(i-1)} & \text{if } i = j, \\ 0.1 & \text{otherwise}; \end{cases} \tag{4.5}$$

and the matrix A to

$$a_{i,j} = \begin{cases} 0 & \text{if } |i-j| > k, \\ 1.0 & \text{if } i = j, \\ 0.07 & \text{otherwise}. \end{cases} \tag{4.6}$$

TABLE 4.3
Time comparison between Tsao's algorithm and ReducekP on the Cray XMP.

n	k	Tsao's algorithm	ReducekP
100	3	0.088	0.009
100	6	0.111	0.026
100	12	0.123	0.064
100	24	0.130	0.155
200	3	0.357	0.025
200	6	0.446	0.064
200	12	0.496	0.151
200	24	0.528	0.374

TABLE 4.4
Numerical results for problem defined by (4.5) *and* (4.6) *for* $k = 2$.

	Relative errors in		Eigenangles	
Eigenvalue	Banded alg.	RSG	Banded alg.	RSG
1.6887E−09	3.45E−04	1.10E−02	5.82E−13	1.87E−11
8.4525E−09	8.01E−04	3.24E−02	6.77E−12	2.74E−10
4.2272E−08	3.46E−04	1.40E−03	1.46E−11	5.92E−11
2.1136E−07	2.33E−04	2.56E−03	4.93E−11	5.40E−10
1.0568E−06	1.15E−04	1.82E−05	1.21E−10	1.92E−11
5.2839E−06	3.06E−05	9.73E−06	1.62E−10	5.14E−11
2.6420E−05	2.39E−06	8.70E−06	6.31E−11	2.30E−10
1.3210E−04	4.96E−06	6.28E−05	6.55E−10	8.29E−09
6.6052E−04	1.66E−06	6.33E−07	1.09E−09	4.18E−10
3.3032E−03	1.55E−06	1.27E−06	5.10E−09	4.17E−09
1.6589E−02	7.71E−07	2.10E−07	1.27E−08	3.45E−09
8.3426E−02	2.03E−08	4.26E−07	1.62E−09	3.40E−08

For $k = 2$ and $n = 12$ the results on the SGI machine are given in Table 4.4 and for $k = 4$ and $n = 12$ the results on the SGI machine are given in Table 4.5. The tables indicate that the algorithm in this section works and they corroborate our previous results for tridiagonal matrices: Algorithm WZ does better on small eigenvalues than the Martin–Wilkinson algorithm, but not as well on large eigenvalues. The true eigenvalues in the tables were computed using the QZ algorithm in double precision.

5. Conclusions. The Wang–Zhao algorithm requires $O(n^3)$ operations and $O(n^2)$ space to solve the symmetric generalized eigenvalue problem on $n \times n$ matrices. We have shown how to modify it so that for $n \times n$ banded matrices of bandwidth $2k + 1$, only $O(n^2k)$ operations and $O(nk)$ space are needed. The new algorithm is a chasing algorithm, which as initially conceived, would involve vectors of length k and would not do well on a vector machine. We have shown that one can do $O(n/k)$ simultaneous chases so that vector lengths can be as large as $O(n/k)$. We have implemented the algorithms to indicate their validity and we have compared them on some ill-conditioned examples

TABLE 4.5
Numerical results for problem defined by (4.5) *and* (4.6) *with* $k = 4$.

	Relative errors in		Eigenangles	
Eigenvalues	Banded alg.	RSG	Banded alg.	RSG
1.6765E−09	2.02E−03	2.96E−02	3.39E−12	4.96E−11
8.3896E−09	1.12E−03	5.76E−02	9.36E−12	4.83E−10
4.1955E−08	3.06E−04	1.22E−04	1.28E−11	5.13E−12
2.0978E−07	8.46E−05	1.45E−02	1.78E−11	3.04E−09
1.0489E−06	8.17E−05	1.30E−04	8.57E−11	1.36E−10
5.2448E−06	2.26E−05	4.18E−04	1.19E−10	2.19E−09
2.6228E−05	4.18E−06	4.42E−06	1.10E−10	1.16E−10
1.3118E−04	6.53E−06	5.45E−06	8.57E−10	7.15E−10
6.5801E−04	4.28E−06	2.07E−06	2.81E−09	1.36E−09
3.3026E−03	4.20E−07	2.58E−09	1.39E−09	8.51E−12
1.6589E−02	2.01E−06	9.09E−09	3.31E−08	1.50E−10
8.3429E−02	3.17E−07	4.91E−08	2.53E−08	3.92E−09

with the Martin–Wilkinson algorithm. The Martin–Wilkinson algorithm usually does worse on small eigenvalues and better on large eigenvalues.

REFERENCES

[1] R. BARTELS AND L. KAUFMAN, *Cholesky factor updating techniques for rank 2 matrix modifications*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 557–592.

[2] C. BISCHOF AND C. VAN LOAN, *The WY representation for products of Householder matrices*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s2–s13.

[3] C. R. CRAWFORD, *Reduction of a band-symmetric generalized eigenvalue problem*, Comm. ACM, 16 (1973), pp. 41–44.

[4] G. H. GOLUB AND V. PEREYRA, *The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate*, SIAM J. Numer. Anal., 10 (1973), pp. 413–432.

[5] G. H. GOLUB AND C. REINSCH, *The singular value decomposition and least squares solution*, Numer. Math., 14 (1970), pp. 403–420.

[6] L. KAUFMAN, *Banded eigenvalue solvers on vector machines*, ACM Trans. Math. Software, 10 (1984), pp. 73–86.

[7] ———, *The generalized Householder transformation and sparse matrices*, Linear Algebra Appl., 90 (1987), pp. 221–234.

[8] R. S. MARTIN AND J. H. WILKINSON, *Reduction of the symmetric eigenvalue problem $Ax = \lambda Bx$ and related problems to standard form*, Numer. Math., 11 (1968), pp. 99–110.

[9] C. B. MOLER AND G. W. STEWART, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241–256.

[10] S. WANG AND S. ZHAO, 1988, personal communication.

[11] ———, *An algorithm for $Ax = \lambda Bx$ with symmetric and positive definite A and B*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 654–660.

[12] B. T. SMITH, J. M. BOYLE, Y. IKEBE, V. C. KLEMA, AND C. B. MOLER, *Matrix Eigensystem Routines: Eispack Guide*, 2nd. Ed., Springer-Verlag, New York, 1976.

[13] G. W. STEWART, *On the perturbation of pseudo-inverses, projections, and linear least squares problems*, SIAM Rev., 19 (1977), pp. 634–662.

[14] G. W. STEWART AND J. SUN, *Matrix Perturbation Theory*, Academic Press, San Diego, 1990.

[15] N. K. TSAO, *On the orthogonal factorization and its updating in band-structured computations*, J. Franklin Inst., 311 (1981), pp. 355–381.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 390-397, April 1993

005

MAXIMUM SUBMATRIX TRACES FOR POSITIVE DEFINITE MATRICES*

I. OLKIN† AND S. T. RACHEV‡

With best wishes to Gene Golub on his 60th birthday.

Abstract. For three p-dimensional jointly distributed random vectors x, y, and z with respective normal marginal distributions $\mathcal{N}(0, \Sigma_{ii})$, $i = 1, 2, 3$, certain covariance matrices are determined that minimize the sum of the L_2-distances of the three vectors. This problem posed in a statistical context is equivalent to maximizing submatrix traces of a positive definite matrix.

Key words. multivariate distributions, covariance matrices, distance of random vectors

AMS(MOS) subject classifications. 15A45, 60E15

1. Introduction. The problem of maximizing submatrix traces of a positive definite matrix has its origins in a statistical context. Suppose that X and Y are jointly distributed p-dimensional random vectors with normally distributed marginals with zero mean vectors and known covariance matrices $\Sigma_{11} > 0$, $\Sigma_{22} > 0$, respectively. ($A \geqq B$ and $A > B$ mean $A - B$ is nonnegative definite, positive definite.) Denote the dispersion matrix of (X, Y) by

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Psi \\ \Psi' & \Sigma_{22} \end{pmatrix}.$$

The L_2-distance between X and Y is

$$E \operatorname{tr} (X - Y)'(X - Y) = \operatorname{tr} (\Sigma_{11} + \Sigma_{22}) - 2 \operatorname{tr} \Psi. \tag{1.1}$$

In a previous paper, Olkin and Pukelsheim (1982) resolved the problem of minimizing (1.1) subject to $\Sigma \geqq 0$, that is,

$$\operatorname*{Max}_{\Sigma \geqq 0} \operatorname{tr} \Psi. \tag{1.2}$$

By translating (1.2) to a dual problem, the solution

$$\Psi_0 = \Sigma_{11}\Sigma_{22}^{1/2}(\Sigma_{22}^{1/2}\Sigma_{11}\Sigma_{22}^{1/2})^{-1/2}\Sigma_{22}^{1/2} \tag{1.3}$$

was obtained. This result was obtained independently by Dowson and Landau (1982) using a different argument. (See also Apitzsch, Fritzsche, and Kirstein (1990).)

This problem suggests a variety of extensions, which are exhibited for the case of three random vectors. Suppose that X, Y, and Z are jointly distributed p-dimensional random vectors with normal marginals $\mathcal{N}(0, \Sigma_{11})$, $\mathcal{N}(0, \Sigma_{22})$, and $\mathcal{N}(0, \Sigma_{33})$, respectively, where $\Sigma_{ii} > 0$, $i = 1, 2, 3$ are known. Denote the joint dispersion matrix by

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Psi_{12} & \Psi_{13} \\ & \Sigma_{22} & \Psi_{23} \\ & & \Sigma_{33} \end{pmatrix}.$$

* Received by the editors February 15, 1991; accepted for publication (in revised form) October 17, 1991.

† Department of Statistics, Stanford University, Stanford, California 94305 (iolkin@playfair.stanford.edu).

‡ Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106 (zarirach@bernoulli.ucsb.edu).

We use Σ_{ij} to denote known covariance matrices and Ψ_{ij} to denote unknown covariance matrices. Two L_2-distances that we consider are

$$\begin{aligned}(1.4)\qquad &E\,\mathrm{tr}\,[(X-Y)'(X-Y)+(X-Z)'(X-Z)+(Y-Z)'(Y-Z)]\\ &\quad=2\,\mathrm{tr}\,(\Sigma_{11}+\Sigma_{22}+\Sigma_{33})-2\,\mathrm{tr}\,(\Psi_{12}+\Psi_{13}+\Psi_{23})\end{aligned}$$

or, with the mean $M=\frac{1}{3}(X+Y+Z)$,

$$\begin{aligned}(1.4')\qquad &E\,\mathrm{tr}\,[(X-M)'(X-M)+(Y-M)'(Y-M)+(Z-M)'(Z-M)]\\ &\quad=\tfrac{2}{3}\,\mathrm{tr}\,(\Sigma_{11}+\Sigma_{22}+\Sigma_{33})-\tfrac{2}{3}\,\mathrm{tr}\,(\Psi_{12}+\Psi_{13}+\Psi_{23}).\end{aligned}$$

Thus both (1.4) and (1.4′) yield the same extremal problems depending on which submatrices of Σ are known. This suggests three cases:

$$(1.5\mathrm{a})\qquad \Sigma_1=\begin{pmatrix}\Sigma_{11} & \Psi_{12} & \Sigma_{13}\\ & \Sigma_{22} & \Sigma_{23}\\ & & \Sigma_{33}\end{pmatrix},\quad \Sigma_{11}>0,\quad \begin{pmatrix}\Sigma_{22} & \Sigma_{23}\\ & \Sigma_{33}\end{pmatrix}>0;$$

$$(1.5\mathrm{b})\qquad \Sigma_2=\begin{pmatrix}\Sigma_{11} & \Psi_{12} & \Psi_{13}\\ & \Sigma_{22} & \Sigma_{23}\\ & & \Sigma_{33}\end{pmatrix},\quad \Sigma_{11}>0,\quad \begin{pmatrix}\Sigma_{22} & \Sigma_{23}\\ & \Sigma_{33}\end{pmatrix}>0;$$

$$(1.5\mathrm{c})\qquad \Sigma_3=\begin{pmatrix}\Sigma_{11} & \Psi_{12} & \Psi_{13}\\ & \Sigma_{22} & \Psi_{23}\\ & & \Sigma_{33}\end{pmatrix},\quad \Sigma_{11}>0,\quad \Sigma_{22}>0,\quad \Sigma_{33}>0.$$

The extremal problems corresponding to each case become

$$(1.6\mathrm{a})\qquad \operatorname*{Max}_{\Sigma_1\geqq 0}\,\mathrm{tr}\,\Psi_{12},$$

$$(1.6\mathrm{b})\qquad \operatorname*{Max}_{\Sigma_2\geqq 0}\,\mathrm{tr}\,(\Psi_{12}+\Psi_{13}),$$

$$(1.6\mathrm{c})\qquad \operatorname*{Max}_{\Sigma_3\geqq 0}\,\mathrm{tr}\,(\Psi_{12}+\Psi_{13}+\Psi_{23}).$$

The proofs provided by Olkin and Pukelsheim (1982) and Dowson and Landau (1982) do not lend themselves to a direct extension for each of the present problems. We first provide an alternative direct proof, which is adaptable to resolving (1.6a) and (1.6b).This is done in § 2. Problem (1.6c) requires a different argument and is discussed in § 3.

2. A direct proof. Our concern in (1.2) is to maximize tr Ψ over the region $\{\Sigma:\Sigma\geqq 0\}$. Since $\Sigma\geqq 0$ if and only if $\Sigma_{22}>0$ (by hypothesis) and $\Sigma_{11}-\Psi\Sigma_{22}^{-1}\Psi'\geqq 0$, we need only consider the latter condition. From the convexity of the set $\{\Psi:\Psi\Sigma_{22}^{-1}\Psi'\leqq\Sigma_{11}\}$ and the linearity of tr Ψ, the extremum will occur at a boundary point

$$(2.1)\qquad \Psi\Sigma_{22}^{-1}\Psi'=\Sigma_{11},$$

which holds if and only if

$$(2.2)\qquad \Psi\Sigma_{22}^{-1/2}=\Sigma_{11}^{1/2}G,$$

where G is orthogonal. (For references and a discussion, see Marshall and Olkin (1979, p. 501).)

Consequently, $\Psi = \Sigma_{11}^{1/2} G \Sigma_{22}^{1/2}$ and the maximum of tr Ψ is tr $G(\Sigma_{22}^{1/2}\Sigma_{11}^{1/2})$. For simplicity of notation, write $B = \Sigma_{22}^{1/2}\Sigma_{11}^{1/2}$, let $\lambda(C)$ denote the eigenvalues of C, and let $\sigma(C) = \lambda^{1/2}(CC')$ denote the singular values. Then

$$\begin{aligned} \operatorname{tr} GB &= \Sigma_1^p \lambda_i(GB) \leqq \Sigma_1^p |\lambda_i(GB)| \leqq \Sigma_1^p \sigma_i(GB) \\ &= \Sigma_1^p \lambda_i^{1/2}(BB') = \Sigma_1^p \lambda_i((BB')^{1/2}) = \operatorname{tr}(BB')^{1/2} \end{aligned} \tag{2.3}$$

(see Marshall and Olkin (1979, p. 232)). Equality in (2.3) is attained for $G = B'(BB')^{-1/2}$. Consequently, the maximum of tr Ψ is achieved at

$$\Psi_0 = \Sigma_{11}^{1/2} G \Sigma_{22}^{1/2} = \Sigma_{11}\Sigma_{22}^{1/2}(\Sigma_{22}^{1/2}\Sigma_{11}\Sigma_{22}^{1/2})^{-1/2}\Sigma_{22}^{1/2},$$

which is (1.3).

We now show that the solutions of (1.6a) and (1.6b) can be accomplished with an argument similar to that given above. In problem (1.6a) we need to maximize tr Ψ_{12} subject to $\Sigma_1 \geqq 0$, which holds provided that

$$\begin{pmatrix} \Sigma_{22} & \Sigma_{23} \\ \Sigma_{32} & \Sigma_{33} \end{pmatrix} > 0$$

and

$$\Sigma_{11} - (\Psi \;\; \Sigma_{13}) \begin{pmatrix} \Lambda_{22} & \Lambda_{23} \\ \Lambda_{32} & \Lambda_{33} \end{pmatrix} \begin{pmatrix} \Psi' \\ \Sigma_{31} \end{pmatrix} \geqq 0, \tag{2.4}$$

where

$$\begin{pmatrix} \Lambda_{22} & \Lambda_{23} \\ \Lambda_{32} & \Lambda_{33} \end{pmatrix} = \begin{pmatrix} \Sigma_{22} & \Sigma_{23} \\ \Sigma_{32} & \Sigma_{33} \end{pmatrix}^{-1}.$$

As before, the extremal occurs at

$$\Sigma_{11} = (\Psi \;\; \Sigma_{13}) \begin{pmatrix} \Lambda_{22} & \Lambda_{23} \\ \Lambda_{32} & \Lambda_{33} \end{pmatrix} \begin{pmatrix} \Psi' \\ \Sigma_{31} \end{pmatrix} = \Psi\Lambda_{22}\Psi' + \Sigma_{13}\Lambda_{32}\Psi' + \Psi\Lambda_{23}\Sigma_{31} + \Sigma_{13}\Lambda_{33}\Sigma_{31}. \tag{2.5}$$

Completing the square in (2.5) yields

$$\begin{aligned} Q &\equiv \Sigma_{11} - \Sigma_{13}\Lambda_{33}\Sigma_{31} + \Sigma_{13}\Lambda_{32}\Lambda_{22}^{-1}\Lambda_{23}\Sigma_{31} \\ &= (\Psi + \Sigma_{13}\Lambda_{32}\Lambda_{22}^{-1})\Lambda_{22}(\Psi + \Sigma_{13}\Lambda_{32}\Lambda_{22}^{-1})'. \end{aligned} \tag{2.6}$$

Note that

$$\Sigma_{22} > 0, \quad \Sigma_{33} > 0, \quad \text{and} \quad \begin{pmatrix} \Sigma_{22} & \Sigma_{23} \\ \Sigma_{32} & \Sigma_{33} \end{pmatrix} > 0,$$

which implies that $\Lambda_{22} = (\Sigma_{22} - \Sigma_{23}\Sigma_{33}^{-1}\Sigma_{32})^{-1} > 0$, so that $Q \geqq 0$. Then (2.6) implies that

$$Q^{1/2}G = (\Psi + \Sigma_{13}\Lambda_{32}\Lambda_{22}^{-1})\Lambda_{22}^{1/2},$$

where G is orthogonal,

$$\Psi = Q^{1/2}G\Lambda_{22}^{-1/2} - \Sigma_{13}\Lambda_{32}\Lambda_{22}^{-1}, \tag{2.7}$$

and

$$\operatorname{tr}\Psi = \operatorname{tr} G(\Lambda_{22}^{-1/2}Q^{1/2}) - \operatorname{tr}\Sigma_{13}\Lambda_{32}\Lambda_{22}^{-1} \leqq \operatorname{tr}(\Lambda_{22}^{-1/2}Q\Lambda_{22}^{-1/2})^{1/2} - \operatorname{tr}\Sigma_{13}\Lambda_{32}\Lambda_{22}^{-1}, \tag{2.8}$$

with equality at $G = (Q^{1/2}\Lambda_{22}^{-1/2})(\Lambda_{22}^{-1/2}Q\Lambda_{22}^{-1/2})^{-1/2}$. Consequently, the minimizing Ψ is

$$\Psi_0 = Q\Lambda_{22}^{-1/2}(\Lambda_{22}^{-1/2}Q\Lambda_{22}^{-1/2})^{-1/2}\Lambda_{22}^{-1/2} - \Sigma_{13}\Lambda_{32}\Lambda_{22}^{-1},$$

where Q is defined by (2.6).

To resolve (1.6b) subject to

$$\Sigma_2 \geqq 0, \quad \Sigma_{11} > 0, \quad \begin{pmatrix} \Sigma_{22} & \Sigma_{23} \\ \Sigma_{32} & \Sigma_{33} \end{pmatrix} > 0,$$

let

$$\Psi = (\Psi_{12}, \Psi_{13}), \qquad \Delta = \begin{pmatrix} \Sigma_{22} & \Sigma_{23} \\ \Sigma_{32} & \Sigma_{33} \end{pmatrix},$$

so that

$$\Sigma_2 = \begin{pmatrix} \Sigma_{11} & \Psi \\ \Psi' & \Delta \end{pmatrix}.$$

Note that $\Sigma_2 \geqq 0$ if and only if $\Sigma_{11} > 0$ and $\Delta \geqq \Psi'\Sigma_{11}^{-1}\Psi$. As before, the extremum occurs at

$$\Sigma_{11}^{1/2}G\Delta^{1/2} = \Psi, \tag{2.9}$$

where now $\Psi : p \times 2p$, and $G : p \times 2p$ with $GG' = I_p$. For simplicity of notation, let

$$C \equiv \Sigma_{11}^{1/2}, \quad \Delta^{1/2} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}, \quad R_{ii} : p \times p, \tag{2.10}$$

and partition $G = (G_1, G_2)$, $G_i : p \times p$. Then (2.9) becomes

$$\Psi = C(G_1, G_2)\begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} = (CG_1R_{11} + CG_2R_{21}, \quad CG_1R_{12} + CG_2R_{22}),$$

and

$$\begin{aligned} \operatorname{tr}(\Psi_{12} + \Psi_{13}) &= \operatorname{tr}[G_1(R_{11} + R_{12})C + G_2(R_{21} + R_{22})C] \\ &= \operatorname{tr}(G_1, G_2)S = \operatorname{tr} GS, \end{aligned} \tag{2.11}$$

where

$$S = \begin{pmatrix} (R_{11} + R_{12})C \\ (R_{21} + R_{22})C \end{pmatrix} : 2p \times p. \tag{2.12}$$

Note that $S' = \Sigma_{11}^{1/2}(I_p, I_p)\Delta^{1/2}$ is of rank p. Following our previous reasoning, the maximum of (2.11) is $\operatorname{tr}(S'S)^{1/2}$ and is achieved at $G = (S'S)^{-1/2}S'$, so that

$$\Psi_0 = \Sigma_{11}^{1/2}(S'S)^{-1/2}S'\Delta^{1/2},$$

where S is defined by (2.12) and (2.10).

3. The general model. Now consider extremal problem (1.6c):

$$\operatorname*{Max}_{\Sigma_3 \geqq 0} \operatorname{tr}(\Psi_{12} + \Psi_{13} + \Psi_{23}). \tag{3.1}$$

The methods of § 2 lead to a prohibitive amount of algebra that is not promising. Instead, we approach this problem afresh.

The triple (x, y, z) of jointly distributed p-dimensional random vectors with respective marginals $\mathcal{N}(0, \Sigma_{11})$, $\mathcal{N}(0, \Sigma_{22})$, $\mathcal{N}(0, \Sigma_{33})$ is called optimal if (3.1) attains its maximum.

Suppose that there exist lower semicontinuous functions f_1, f_2, f_3 such that

$$x \in \partial f_3(z), \quad y \in \partial f_1(x), \quad z \in \partial f_2(y), \quad \text{a.s.,} \tag{3.2}$$

where $\partial f(x)$ denotes the subdifferential of f in x. Then (x, y, z) is optimal. In fact, for any other triple $(\tilde{x}, \tilde{y}, \tilde{z})$ with the same marginal distributions as (x, y, z), let

$$\Delta(x, y, z) = \langle x, y \rangle + \langle x, z \rangle + \langle y, z \rangle,$$

where $\langle \cdot, \cdot \rangle$ is an inner product in $\mathbb{R}^p$. Then

$$\begin{aligned} E\Delta(\tilde{x}, \tilde{y}, \tilde{z}) &\leqq E\{f_1(\tilde{x}) + f_1^*(\tilde{y}) + f_2(\tilde{y}) + f_2^*(\tilde{z}) + f_3(\tilde{z}) + f_3^*(\tilde{x})\} \\ &= E\{f_1(x) + f_1^*(y) + f_2(y) + f_2^*(z) + f_3(z) + f_3^*(x)\} \\ &= E\Delta(x, y, z), \end{aligned}$$

where $f_j^*(y) = \sup_x \{\langle x, y \rangle - f_j(x)\}$ is the conjugate of f_j. Consequently,

$$f_i(x) + f_i^*(y) \geqq \langle x, y \rangle$$

for all $x, y \in \mathbb{R}^p$, and

$$f_i(x) + f_i^*(y) = \langle x, y \rangle$$

if and only if $y \in \partial f_i(x)$.

For the special univariate case $p = 1$, and for $F_j \sim \mathcal{N}(0, \sigma_j^2)$, $j = 1, 2, 3$, $\{x \equiv F_1^{-1}(u), y \equiv F_2^{-1}(u), z \equiv F_3^{-1}(u)\}$, u uniform on $[0, 1]$, is an optimal triple. (See Lorentz (1953) and Tchen (1980).) To see this, take $f_j(x) = \int_0^x F_{j+1}^{-1} \circ F_j(u) du$, $j = 1, 2, 3$ (with $F_4 \equiv F_1$); then (3.2) holds.

Remark. For other examples, see also Rüschendorf and Rachev (1990). Although the case when $p = 1$ yields an optimal triple and (3.2) holds, this condition does not seem to be necessary for optimality. This is in contrast to some of the special cases considered in Rüschendorf and Rachev (1990).

Remark. Suppose the probabilities μ_j, $j = 1, 2, 3$ on $\mathbb{R}^p$ are of the form

$$\mu_2 = \mu_1 \circ T_1^{-1}, \quad \mu_3 = \mu_2 \circ T_2^{-1}, \quad \mu_1 = \mu_3 \circ T_3^{-1}, \tag{3.3}$$

where T_j are symmetric and positive semidefinite, and

$$T_3 T_2 T_1 = I. \tag{3.4}$$

By (3.2), if x is μ_1-distributed, $y = T_1 x$, and $z = T_2 y$, then (x, y, z) is optimal.

In the particular example where the marginals are normal,

$$\begin{aligned} T_1 &= T(\Sigma_{11}, \Sigma_{22}) \equiv \Sigma_{22}^{1/2}(\Sigma_{22}^{1/2}\Sigma_{11}\Sigma_{22}^{1/2})^{-1/2}\Sigma_{22}^{1/2}, \\ T_2 &= T(\Sigma_{22}, \Sigma_{33}), \\ T_3 &= T(\Sigma_{33}, \Sigma_{11}). \end{aligned}$$

Assuming that (3.4) holds, then the triple (x, T_1x, T_2T_1x) is optimal, and the maximum of (3.1) is

$$E\{\langle x, T_1x\rangle + \langle y, T_2x\rangle + \langle z, T_3x\rangle\}$$
$$= \operatorname{tr}(\Sigma_{11}^{1/2}\Sigma_{22}\Sigma_{11}^{1/2})^{1/2} + \operatorname{tr}(\Sigma_{22}^{1/2}\Sigma_{33}\Sigma_{22}^{1/2})^{1/2} + \operatorname{tr}(\Sigma_{33}^{1/2}\Sigma_{11}\Sigma_{33}^{1/2})^{1/2}.$$

Remark. The condition (3.4) may appear unusual. However, it is satisfied if Σ_{11}, Σ_{22}, Σ_{33} commute. In this case Σ_{jj}, $j = 1, 2, 3$ can be simultaneously diagonalized by the same orthogonal matrix, that is,

$$\Sigma_{jj} = \Gamma D_j \Gamma',$$

where D_j is a diagonal matrix and Γ is orthogonal. Then

$$T_1 = \Gamma D_2^{1/2} D_1^{-1/2}\Gamma',$$
$$T_2 = \Gamma D_3^{1/2} D_2^{-1/2}\Gamma',$$
$$T_3 = \Gamma D_1^{1/2} D_3^{-1/2}\Gamma',$$

so that $T_3T_2T_1 = I$. Of course, the simplest example in which Σ_{11}, Σ_{22}, and Σ_{33} commute is the case where $\Sigma_{11} = \Sigma_{22} = \Sigma_{33}$. Then $T_1 = T_2 = T_3 = I$. They also commute in other circumstances, such as when the elements of x, of y, of z are exchangeable random variables.

4. Historical connections with linear programming. An extension of the well-known transportation problem in linear programming to an infinite set of origins (destinations) was considered very early on by Monge (1781). The formulation for infinite-dimensional transportation problems was initiated by Kantorovich (1942), (1948). In particular, the dual form of (1.2) is

$$\text{(4.1)} \qquad \operatorname*{Max}_{\Omega}\left\{\int f\, d\mu_1 + \int g\, d\mu_2\right\},$$

where $f : \mathbb{R}^p \to R$, $g : \mathbb{R}^p \to R$, $\mu_i \sim \mathcal{N}(0, \Sigma_{ii})$, $i = 1, 2$;

$$\text{(4.2)} \qquad \begin{aligned}\Omega = \{f, g : f, g \text{ bounded}, \|f\|_L < \infty, \|g\|_L < \infty,\\ f(x) + g(y) \leqq \|x - y\|_2 \text{ for all } x, y \in \mathbb{R}^p\},\end{aligned}$$

where

$$\|f\|_L \equiv \sup_{x \neq y} \frac{\|f(x) - f(y)\|}{\|x - y\|_2}.$$

The explicit solution for the unidimensional case ($p = 1$) was obtained by Gini (1914), Hoeffding (1940), and Fréchet (1957). An explicit solution in a more general context:

$$\text{(4.3)} \qquad L_r(\mu, \nu) = \operatorname{Min}\{E\|X - Y\|_r^r, X \sim \mu, Y \sim \nu\},$$

where $\|X\|_r = (\sum_1^{253p} |x_i|^r)^{1/r}$, $r \geqq 1$, and μ, ν are arbitrary laws on $\mathbb{R}^p$, $p > 1$, is a well-known open problem. See, for example, the discussion in Dobrushin (1970) about the relationship between (4.3), with $r = 1$, and the problem of uniqueness of the Gibbs field. The solution of (4.3) with $r = 2$ for μ, ν normal was obtained in Olkin and Pukelsheim (1982). Dowson and Landau (1982) gave an alternative proof. (This proof has a gap; cf. Givens and Shortt (1984) and Gelbrich (1988).) Alternative solutions and some extensions are given in Knott and Smith (1984); Smith and Knott (1987), (1990); Rachev and Rüschendorf (1991); Rüschendorf and Rachev (1990).

The problem of finding explicit solutions of (1.4) or (1.6c) for $p = 1$ dates from the classical paper of Lorentz (1953); see also Tchen (1980). Dual representations for (1.2) and (4.3) are given in Kellerer (1984) and Rachev (1984). Note that L_2 in (4.3) has nice topological properties; it metrizes the weak convergence plus the convergence of the second moments $\int \|X\|_2^2 \mu(dx)$. (See Mallows (1972), Bickel and Freedman (1981), de Acosta (1982), and Rachev (1983).) Analogously, if

$$L_2(\mu_1, \mu_2, \mu_3) \equiv \text{Min}\,\{E(\|X-Y\|_2^2 + \|X-Z\|_2^2 + \|Y-Z\|_2^2)\}, \tag{4.4}$$

where $X \sim \mu_1$, $Y \sim \mu_2$, and $Z \sim \mu_3$, the convergence

$$L_2(\mu_1^{(n)}, \mu_2^{(n)}, \mu_3^{(n)}) \to 0 \tag{4.5}$$

is equivalent to a "merging" of sequences $\mu_1^{(n)}$, $\mu_2^{(n)}$, and $\mu_3^{(n)}$ in L_2-metric. That is, (4.5) is equivalent to

$$L_2(\mu_1^{(n)}, \mu_2^{(n)}) \to 0, \quad L_2(\mu_1^{(n)}, \mu_3^{(n)}) \to 0, \quad L_2(\mu_2^{(n)}, \mu_3^{(n)}) \to 0. \tag{4.6}$$

The notion of merging of two sequences of probabilities was studied in Diaconis and Freedman (1984), D'Aristotile, Diaconis, and Freedman (1988), and Dudley (1989, § 11.7). Using the equivalence of (4.5) and (4.6) permits an extension of their results to the merging of a family of sequences of probabilities.

REFERENCES

W. Apitzsch, B. Fritzsche, and B. Kirstein (1990), *A Schur analysis approach to minimum distance problems*, Linear Algebra Appl., 141, pp. 105–122.

P. J. Bickel and D. A. Freedman (1981), *Some asymptotic theory for the bootstrap*, Ann. Statist., 9, pp. 1196–1217.

A. D'Aristotile, P. Diaconis, and D. A. Freedman (1988), *On merging of probabilities*, Sankhyā, Ser. A, 50, pp. 363–380.

A. de Acosta (1982), *Invariance principles in probability for triangle arrays of B-valued random vectors and some applications*, Ann. Probab., 10, pp. 346–373.

P. Diaconis and D. A. Freedman (1984), *A note on weak star uniformities*, Tech. Rep. 39, Dept. of Statistics, Univ. of California, Berkeley, CA.

R. L. Dobrushin (1970), *Prescribing a system of random variables by conditional distributions*, Theory Probab. Appl., 15, pp. 458–486.

D. C. Dowson and B. V. Landau (1982), *The Fréchet distance between multivariate normal distributions*, J. Multivariate Anal., 12, pp. 450–455.

R. M. Dudley (1989), *Real Analysis and Probability*, Brooks/Cole Publishing Co., Pacific Grove, CA.

M. Fréchet (1957), *Sur la distance de deux lois de probabilité*, C. R. Acad. Sci. Paris, 244, pp. 689–692.

M. Gelbrich (1988), *On a formula for the L^p Wasserstein metric between measures on Euclidean and Hilbert spaces*, Preprint, Sektion Mathematic der Humboldt-Universitat zu Berlin, No. 179, Berlin.

C. Gini (1914), *Di una misura della relazioni tra le graduatorie di due caratteri. L'indice di cograduazione.* (Appendix to Le Elezioni Generale Politiche del 1913 nel Commune di Roma, A. Mancini, Ludovico Cecchini, Rome.)

C. R. Givens and R. M. Shortt (1984), *A class of Wasserstein metrics for probability distributions*, Michigan Math. J., 31, pp. 231–240.

W. Hoeffding (1940), *Maszstabinvariante Korrelationstheorie*, Univ. of Berlin Schriften Math. Angew. Math., 5, pp. 181–233.

L. V. Kantorovich (1942), *On transfer of masses*, Dokl. Akad. Nauk. SSSR, 37, pp. 7, 8, and 227–239. (In Russian.)

——— (1948), *On a problem of Monge*, Uspekhi Mat. Nauk 3, 2, pp. 225–226. (In Russian.)

H. G. Kellerer (1984), *Duality theorems for marginal problems*, Z. Wahrsch. Verw. Gebiete, 67, pp. 299–432.

M. Knott and C. S. Smith (1984), *On the mapping of distributions*, J. Optim. Theory Appl., 43, pp. 39–49.

G. G. Lorentz (1953), *An inequality for rearrangements,* Amer. Math. Monthly, 60, pp. 176–179.

C. L. Mallows (1972), *A note on asymptotic joint normality*, Ann. Math. Statist., 43, pp. 508–515.

A. W. Marshall and I. Olkin (1979), *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.

G. Monge (1781), *Mémoire sur la théorie des déblais et des remblais*, Mém. Math. Phys. Acad. Roy. Sci., Paris, pp. 667–709.

I. Olkin and F. Pukelsheim (1982), *The distance between two random vectors with given dispersion matrices*, Linear Algebra Appl., 48, pp. 257–263.

S. T. Rachev (1983), *Minimal metrics in the random variable space*, Pub. Inst. Statist., Univ. Paris, XXVII, pp. 27–47.

——— (1984), *The Monge–Kantorovich mass transference problem and its stochastic applications*, Theory Probab. Appl., 24, pp. 647–671.

S. T. Rachev and L. Rüschendorf (1991), *Recent results in the theory of probability metrics*. Statist. Decisions, 9, pp. 327–274.

L. Rüschendorf and S. T. Rachev (1990), *A characterization of random variables with minimum L^2-distance*, J. Multivariate Anal., 32, pp. 48–59.

C. S. Smith and M. Knott (1987), *Note on the optimal transportation of distributions*, J. Optim. Theory Appl., 52, pp. 323–329.

——— (1990), *C-cyclic monotone relations*, Preprint, Dept. of Statistical and Mathematical Sciences, London School of Economics and Political Science, London, U.K.

A. H. Tchen (1980), *Inequalities for distributions with given marginals*, Ann. Probab., 8, pp. 814–827.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 398–407, April 1993

006

SIGN CONTROLLABILITY OF A NONNEGATIVE MATRIX AND A POSITIVE VECTOR*

CHARLES R. JOHNSON†, VOLKER MEHRMANN‡, AND D. DALE OLESKY§

Abstract. An n-by-n matrix A and a vector b are controllable if and only if the matrix $[b, Ab, A^2b, \ldots, A^{n-1}b]$ has rank n. An array with each entry equal to +, −, or 0 is a sign pattern. If **A** and **B** are sign patterns of orders n-by-n and n-by-1, respectively, then the pair $(\mathbf{A}, \mathbf{B})$ is called sign controllable if (A, b) is controllable for all $A \in \mathbf{A}$ and for all $b \in \mathbf{B}$. Sign controllability of a nonnegative sign pattern **A** and a positive sign pattern **B** are characterized, and sufficient conditions for other cases of the sign controllability problem are given.

Key words. controllability, sign controllability, nonnegative matrix

AMS(MOS) subject classifications. 15A99, 93B05

1. Introduction. Let $A \equiv (a_{ij})$ be an n-by-n real matrix and let $b \equiv (b_i)$ denote a vector with n real entries. The pair (A, b) is said to be *controllable* if and only if the n-by-n matrix $C(A, b) = [b, Ab, A^2b, \ldots, A^{n-1}b]$ has rank n. Many equivalent conditions for controllability are known (see [7]); for example, (A, b) is controllable if and only if the n-by-$(n+1)$ matrix $D(A, b, \lambda) = [A - \lambda I, b]$ has rank n for all complex scalars λ.

Let **A** denote an n-by-n *sign pattern*, that is, an n-by-n array with each entry equal to +, −, or 0. For example,

$$\mathbf{A} = \begin{bmatrix} + & 0 & - \\ - & + & + \\ 0 & + & 0 \end{bmatrix}.$$

We write $A \in \mathbf{A}$ if A is an n-by-n matrix with sgn (a_{ij}) equal to the (i, j) entry of **A**. The n-by-1 *sign pattern* is denoted by **B**.

We consider the following qualitative problem: Determine the set of all sign patterns **A** and **B** such that if $A \in \mathbf{A}$ and $b \in \mathbf{B}$, then (A, b) is controllable. Such sign patterns **A** and **B** are said to *require* the property of controllability, and we say that $(\mathbf{A}, \mathbf{B})$ is *sign controllable*. See [3] for a definition and discussion of "require" problems of combinatorial matrix analysis.

If all matrices with a fixed sign pattern are nonsingular, then each such matrix is said to be *sign nonsingular*; see, for example, [1] and [4]. If $C(A, b)$ is sign nonsingular for all $A \in \mathbf{A}$ and $b \in \mathbf{B}$, then clearly $(\mathbf{A}, \mathbf{B})$ is sign controllable. However, as the following example shows, the converse is not true.

Example 1. Let

$$\mathbf{A} = \begin{bmatrix} 0 & + & 0 \\ 0 & + & + \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} + \\ + \\ + \end{bmatrix}.$$

* Received by the editors January 21, 1991; accepted for publication (in revised form) August 2, 1991.

† Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23185 (crjohn@wmvm1.bitnet).

‡ Fakultät für Mathematik, Universität Bielefeld, Bielefeld, Germany. This author's research was performed while he was visiting the Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, D-5100 Aachen, Germany (mehrmann@math1.mathematik.uni-bielefeld.de).

§ Department of Computer Science, University of Victoria, Victoria, British Columbia V8W 3P6, Canada(dolesky@uvunix.uvic.ca). The work of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A-8214 and the University of Victoria President's Committee on Faculty Research and Travel.

If $A \in \mathbf{A}$ and $b \in \mathbf{B}$, then

$$\det C(A, b) = -a_{12}a_{23}b_3^2(a_{22}b_2 + a_{23}b_3) < 0,$$

implying that $(\mathbf{A}, \mathbf{B})$ is sign controllable. However, for all $A \in \mathbf{A}$ and $b \in \mathbf{B}$, the sign pattern of $C(A, b)$ is

$$\begin{bmatrix} + & + & + \\ + & + & + \\ + & 0 & 0 \end{bmatrix},$$

which is not a possible sign pattern of a sign nonsingular matrix.

The problem of characterizing sign controllability has important application in the design of control problems. Typically the parameters that determine relationships in a linear system $\dot{x} = Ax + bu$ describing a control system are subject to measurement or modelling errors. So, when designing such a control system, it is of interest to know whether or not such a system is controllable regardless of these errors; see, for example [2] or [6]. A natural question in this context is whether or not controllability depends solely on the signs of the entries of A and b. In general, this problem seems difficult. Here, we consider the special case in which the entries of A are nonnegative and the entries of b are positive, which, for example, is a reasonable assumption in economic and biological models. In §§ 2 and 3 of this paper, $\mathbf{A}$ denotes an n-by-n *nonnegative sign pattern* and $\mathbf{B}$ denotes the n-by-1 *positive sign pattern*. If $A \in \mathbf{A}$ and $b \in \mathbf{B}$, we write $A \geqq 0$ and $b > 0$.

Section 2 contains some notation, definitions, and lemmas that are used in § 3 to prove our main result, a characterization of sign controllability of a nonnegative matrix and a positive vector. More general cases of sign controllability are considered in § 4.

2. Preliminary results. Throughout this paper we let u_k denote the vector of k entries, each of which is equal to 1, and let I_k denote the k-by-k identity matrix. For completeness, we state the following definitions. A square matrix A is *irreducible* if and only if there does not exist a permutation matrix P such that

$$PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

where A_{11} and A_{22} are square, nonempty submatrices. For every n-by-n matrix A, there exists a permutation matrix P such that

$$PAP^T = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ & A_{22} & \cdots & A_{2k} \\ 0 & & \ddots & \vdots \\ & & & A_{kk} \end{bmatrix},$$

where $1 \leqq k \leqq n$ and each submatrix A_{ii}, $1 \leqq i \leqq k$, is either 1-by-1 or irreducible. This is called the *Frobenius normal form* of A. Finally, an n-by-n matrix $A \geqq 0$ is *row stochastic* if $Au_n = u_n$. Note that if A is row stochastic, so is A^k for all $k \geqq 1$.

We now state two lemmas, which are used to prove Theorem 1 in § 3.

LEMMA 1. *Let $b > 0$ be a vector with m entries and let $R \geqq 0$ be an upper triangular matrix of order m with exactly one positive diagonal entry, say r_{kk} $(1 \leqq k \leqq m)$. Then there exists a nonsingular matrix $L \geqq 0$ such that $L^{-1}RL$ is upper triangular with exactly one nonzero diagonal entry, which is in the* (1, 1) *position and is equal to r_{kk}. Moreover, the first entry of $L^{-1}b$ is positive.*

Proof. Without loss of generality, we may assume that $m \geqq 2$ and $k = m$, since otherwise we can restrict consideration to the leading principal submatrix of R of order k and can take L in the form

$$\begin{bmatrix} L_{11} & 0 \\ 0 & I_{m-k} \end{bmatrix},$$

where L_{11} is of order k. Also, without loss of generality, we may assume that no row of R is all zeros, since otherwise we can find a permutation matrix Q such that

$$Q^T R Q = \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix},$$

where R_{11} is k-by-k upper triangular with no rows that are all zero, and, in this case, we can determine L_{11} such that $L_{11}^{-1} R_{11} L_{11}$ has the required form, and let

$$L = Q \begin{bmatrix} L_{11} & 0 \\ 0 & I_{m-k} \end{bmatrix}.$$

We specify a set of $m - 1$ similarity transformations that moves r_{mm} to the $(1, 1)$ position, each transformation moving the entry up one diagonal position. Since $r_{m-1,m}$ is necessarily nonzero (as $r_{m-1,m-1} = 0$), we can define

$$L_m = \left[\begin{array}{c|cc} I_{m-2} & \multicolumn{2}{c}{0} \\ \hline 0 & 1 & 0 \\ & c_m & 1 \end{array}\right],$$

where $c_m = r_{mm}/r_{m-1,m}$. Then

$$R_m \equiv L_m^{-1} R L_m$$

$$= \begin{bmatrix} 0 & r_{12} & r_{13} & \cdots & r_{1,m-2} & r_{1,m-1} + c_m r_{1m} & r_{1m} \\ & 0 & r_{23} & \cdots & r_{2,m-2} & r_{2,m-1} + c_m r_{2m} & r_{2m} \\ & & \ddots & \ddots & \vdots & \vdots & \vdots \\ & & & \ddots & r_{m-3,m-2} & \vdots & \vdots \\ & & & & 0 & r_{m-2,m-1} + c_m r_{m-2,m} & r_{m-2,m} \\ & 0 & & & 0 & r_{mm} & r_{m-1,m} \\ & & & & 0 & 0 & 0 \end{bmatrix}.$$

The above step can be repeated for $j = m - 1, m - 2, \ldots, 2$ by defining

$$L_j = \left[\begin{array}{c|cc|c} I_{j-2} & \multicolumn{2}{c|}{0} & 0 \\ \hline 0 & 1 & 0 & 0 \\ & c_j & 1 & \\ \hline 0 & \multicolumn{2}{c|}{0} & I_{m-j} \end{array}\right],$$

where c_j is equal to r_{mm} divided by the $(j - 1, j)$ entry of R_{j+1}. (Note that the assumption that no row of R is all zeros ensures that the $(j - 1, j)$ entry of R_{j+1} is nonzero since this entry is equal to

$$r_{j-1,j} + \sum_{i=j+1}^{m} c_i r_{j-1,i},$$

where all $c_i > 0$ and at least one entry $r_{j-1,i} > 0$.) Letting $L = L_m L_{m-1} \cdots L_2$, then $L^{-1}RL$ is as required. Furthermore, for $2 \leqq j \leqq m$, the first $j-1$ entries of $L_j^{-1} \cdots L_{m-1}^{-1} L_m^{-1} b$ are equal to $(b_1, b_2, \ldots, b_{j-1})^T$, so that the first entry of $L^{-1}b = L_2^{-1} \cdots L_{m-1}^{-1} L_m^{-1} b$ is equal to b_1 and thus is positive. □

LEMMA 2. *Let* **A** *denote an n-by-n nonnegative sign pattern, let* **B** *denote the n-by-1 positive sign pattern, let $A \in$* **A** *and $b \in$* **B**. *If there exists a permutation matrix P such that*

$$PAP^T = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

where A_{11} is either a nonzero scalar or is an irreducible submatrix of order $\geqq 2$, and A_{22} is upper triangular with exactly one nonzero diagonal element, then (**A**, **B**) *is not sign controllable.*

Proof. We determine a specific (numeric) matrix A and vector b such that (A, b) is not controllable. Let k_1 denote the order of A_{11} and let $b = (b_1, b_2)^T$, where b_1 has k_1 entries. Let the nonzero diagonal entry of A_{22} equal 1. For any choice of the other entries of A_{22}, by Lemma 1 there exists a nonsingular matrix $L \geqq 0$ such that $L^{-1}A_{22}L$ is upper triangular with exactly one nonzero diagonal entry, namely its (1, 1) entry which is equal to 1. Also, the first entry of $L^{-1}b_2$ is positive. Let

$$\tilde{A} = \begin{bmatrix} I & 0 \\ 0 & L^{-1} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & L \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12}L \\ 0 & L^{-1}A_{22}L \end{bmatrix}$$

and

$$\tilde{b} = \begin{bmatrix} I & 0 \\ 0 & L^{-1} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ L^{-1}b_2 \end{bmatrix},$$

and partition these, respectively, as

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ 0 & \hat{A}_{22} \end{bmatrix} \quad \text{and} \quad \hat{b} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix},$$

where $\hat{A}_{11}$ is of order $k_1 + 1$ and $\hat{b}_1$ has $k_1 + 1$ entries. Let $k_2 = n - k_1 - 1$ denote the order of $\hat{A}_{22}$ and the size of $\hat{b}_2$. Then

$$\hat{A}_{11} = \begin{bmatrix} A_{11} & a_1 \\ 0 & 1 \end{bmatrix},$$

where A_{11} is as before and a_1 is equal to the first column of $A_{12}L$ (and is nonnegative for any choice of the entries of A_{12}). Also, $\hat{A}_{22}$ is strictly upper triangular, $\hat{b}_1 > 0$, and $\hat{b}_2$ consists of the last k_2 entries of $L^{-1}b_2$. Consider the controllability matrix $C(\hat{A}, \hat{b})$, which is equal to

$$\left[\begin{array}{c|c|c|c|c|c|c|c} \hat{b}_1 & \hat{A}_{11}\hat{b}_1 + \hat{A}_{12}\hat{b}_2 & \cdots & \hat{A}_{11}^{k_2-1}\hat{b}_1 + F\hat{b}_2 & \hat{A}_{11}^{k_2}\hat{b}_1 + G\hat{b}_2 & \hat{A}_{11}g & \cdots & \hat{A}_{11}^{k_1}g \\ \hat{b}_2 & \hat{A}_{22}\hat{b}_2 & & \hat{A}_{22}^{k_2-1}\hat{b}_2 & 0 & 0 & & 0 \end{array}\right],$$

where F and G are matrices of order $(k_1 + 1)$-by-k_2 and are a sum of products of $\hat{A}_{11}$, $\hat{A}_{12}$, and $\hat{A}_{22}$, and $g = \hat{A}_{11}^{k_2}\hat{b}_1 + G\hat{b}_2$. As A_{11} is either a nonzero scalar or is irreducible of order $\geqq 2$, and as the entries of a_1 are nonnegative linear combinations of the entries of A_{12}, the entries of $\hat{A}_{11}$ can be chosen so that it is row stochastic and nonsingular. (For example, choose the positive entries of A_{12} to be arbitrarily small, and choose all of the positive entries of A_{11} to be arbitrarily small except for those on one transversal of A_{11}, which can then be chosen to make $\hat{A}_{11}$ row stochastic.) By choosing b_2 sufficiently small, $\hat{b}_2$ and thus $G\hat{b}_2$ can be made arbitrarily close to the 0 vector. Define $\hat{b}_1 = (\hat{A}_{11}^{k_2})^{-1}(u_{k_1+1} - G\hat{b}_2)$; this vector can be made arbitrarily close to u_{k_1+1}, and thus made positive. With this choice of $\hat{b}_1$, we obtain $g = u_{k_1+1}$. Hence $C(\hat{A}, \hat{b})$ has $k_1 + 1 \geqq 2$ identical columns of the form $(u_{k_1+1}, 0)^T$, since $\hat{A}_{11}$ is row stochastic. Thus $(\hat{A}, \hat{b})$ is not controllable. Because (A, b) is controllable if and only if $(\hat{A}, \hat{b})$ is controllable (see, for example, [5]), $(\mathbf{A}, \mathbf{B})$ is not sign controllable. □

3. The main result. The following theorem characterizes sign controllability of a nonnegative matrix and a positive vector.

THEOREM 1. *Let $n \geqq 2$, let $\mathbf{A}$ denote an n-by-n nonnegative sign pattern and let $\mathbf{B}$ denote the n-by-1 positive sign pattern. Then $(\mathbf{A}, \mathbf{B})$ is sign controllable if and only if*

(i) rank $\hat{A} = n - 1$ *for all $\hat{A} \in \mathbf{A}$, and for each $\hat{A} \in \mathbf{A}$, there exists a permutation matrix P such that $A = P\hat{A}P^T$ has the following properties*:

(ii) $a_{nj} = 0$ *for* $1 \leqq j \leqq n$; *and*

(iii) *A is upper triangular with at most one positive diagonal entry.*

Proof of sufficiency. Let $\hat{A} \in \mathbf{A}$ and $b \in \mathbf{B}$. Because $(\hat{A}, b)$ is controllable if and only if $(P\hat{A}P^T, Pb)$ is controllable (see [5]), without loss of generality, suppose that $A = P\hat{A}P^T$ is upper triangular with at most one positive diagonal entry, say $a_{kk} = p \geqq 0$, where $1 \leqq k \leqq n - 1$. Thus both (ii) and (iii) are satisfied by A, and $D(A, Pb, \lambda) \equiv [A - \lambda I, Pb]$ has the form

$$\begin{bmatrix}
-\lambda & * & * & \cdots & * & * & * & \cdots & * & + \\
 & -\lambda & * & \cdots & * & * & * & \cdots & * & + \\
 & & \ddots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
 & & & -\lambda & * & * & * & \cdots & * & + \\
 & & & & p-\lambda & * & * & \cdots & * & + \\
 & & & & & -\lambda & * & \cdots & * & + \\
 0 & & & & & & \ddots & & \vdots & \vdots \\
 & & & & & & & \ddots & \vdots & \vdots \\
 & & & & & & & & -\lambda & +
\end{bmatrix},$$

where $*$ denotes a nonnegative entry.

If $\lambda = 0$, then this matrix has rank n since A has rank $n - 1$ (by (i)) and Pb is linearly independent of the column vectors of A (as it has a "+" in the nth position).

If $\lambda \neq 0$ and $\lambda \neq p$, then $A - \lambda I$ is upper triangular with a nonzero diagonal, and so has rank n.

If $\lambda = p$ (and $p > 0$), then $D(A, Pb, \lambda)$ has the form

$$\left[\begin{array}{ccccc|c|cccccc}
-\lambda & * & * & \cdots & * & * & * & \cdots & \cdots & & * & + \\
 & -\lambda & * & \cdots & * & * & * & \cdots & \cdots & & * & + \\
 & & \ddots & & \vdots & \vdots & \vdots & & & & \vdots & \vdots \\
 & 0 & & \ddots & * & * & * & \cdots & \cdots & & * & + \\
 & & & & -\lambda & * & * & \cdots & \cdots & & * & + \\
\hline
 & & & & & 0 & * & * & * & \cdots & * & + \\
 & & & & & 0 & -\lambda & * & * & \cdots & * & + \\
 & & & & & \vdots & & -\lambda & * & \cdots & * & + \\
 & 0 & & & & \vdots & & & \ddots & \ddots & \vdots & \vdots \\
 & & & & & 0 & 0 & & & \ddots & * & + \\
 & & & & & 0 & & & & & -\lambda & +
\end{array}\right] \equiv \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & 0 & A_{23} \end{bmatrix},$$

where each $*$ is a nonnegative entry. This matrix has rank n since A_{11} is nonsingular and A_{23} is sign nonsingular (see, for example, [1]).

Thus $D(A, Pb, \lambda)$ has rank n for all $\lambda \in \mathbf{C}$, proving sufficiency.

Note. If A has either no positive diagonal entry or if the $(1, 1)$ entry of A is positive, then A has the form

$$\begin{bmatrix} a_{11} & a_{12} & * & * & * & \cdots & * \\ & 0 & + & * & * & \cdots & * \\ & & 0 & + & * & \cdots & * \\ & & & \ddots & \ddots & \ddots & \vdots \\ & & & & \ddots & \ddots & * \\ & 0 & & & & \ddots & + \\ & & & & & & 0 \end{bmatrix},$$

where at least one of a_{11} and a_{12} is positive and each $*$ (without restriction) may be 0 or $+$. In this case, $C(A, Pb)$ has the form

$$\begin{bmatrix} + & + & + & \cdots & + & + \\ + & + & + & \cdots & + & \\ \vdots & \vdots & \vdots & \iddots & & \\ + & + & + & & & \\ + & + & & & 0 & \\ + & & & & & \end{bmatrix},$$

which is clearly of rank n, implying that (A, Pb) is controllable.

Proof of necessity. Assume that $(\mathbf{A}, \mathbf{B})$ is sign controllable.

If $\mathbf{A}$ does not have a row consisting entirely of zeros, then we can choose $\hat{A} \in \mathbf{A}$ to be row stochastic and can let $b = u_n$. Then $C(\hat{A}, b)$ is singular. Thus $\mathbf{A}$ must have a row consisting entirely of zeros, and (ii) holds.

Since $D(\hat{A}, b, \lambda)$ has rank n for all $\hat{A} \in \mathbf{A}$ and for all $\lambda \in \mathbf{C}$, the case $\lambda = 0$ implies that rank $\hat{A} \geqq n - 1$. As $\hat{A}$ must have a row consisting entirely of zeros (by (ii)), condition (i) holds.

To prove (iii), assume without loss of generality that each $A \in \mathbf{A}$ is in Frobenius normal form

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ & A_{22} & \cdots & A_{2m} \\ 0 & & \ddots & \vdots \\ & & & A_{mm} \end{bmatrix},$$

where each submatrix A_{ii} is of order 1 or is irreducible and of order $\geqq 2$. Partition $b = (b_1, b_2, \ldots, b_m)^T$ conformally with A. Our proof is by induction on the order of A (or $\mathbf{A}$). We now assume that for all nonnegative sign patterns $\mathbf{A}$ of order $r \leqq n - 1$ and the positive sign pattern $\mathbf{B}$ of size r, if $(\mathbf{A}, \mathbf{B})$ is sign controllable and if $\hat{A} \in \mathbf{A}$, then there exists a permutation matrix P such that $A = P\hat{A}P^T$ satisfies (iii).

Assume that $(\mathbf{A}, \mathbf{B})$ is sign controllable, where $\mathbf{A}$ is n-by-n. The following cases exhaust all possible forms for an n-by-n matrix $A \in \mathbf{A}$.

Case a. If $m = 1$, then $A = A_{11}$ is of order $n \geqq 2$ and A does not satisfy (iii). However, A_{11} may be chosen to be row stochastic. For $b = u_n$, $C(A, b)$ is singular and thus (A, b) is not controllable.

If $m \geqq 2$ (Cases b and c below), then partition A and b, respectively, as

$$\begin{bmatrix} A_{11} & \tilde{A}_{12} \\ 0 & \tilde{A}_{22} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} b_1 \\ \tilde{b}_2 \end{bmatrix}.$$

Since (A, b) controllable implies that $(\tilde{A}_{22}, \tilde{b}_2)$ is controllable, it follows from the inductive hypothesis that $\tilde{A}_{22}$ satisfies (iii). Let k_2 denote the order of $\tilde{A}_{22}$.

Case b. If $\tilde{A}_{22}$ is strictly upper triangular and A_{11} is irreducible and of order $k_1 \geqq 2$, then the entries of A_{11} may be chosen so that it is nonsingular and row stochastic. In this case,

$$C(A, b) = \left[\begin{array}{c|c|c|c|c|c|c} b_1 & A_{11}b_1 + \tilde{A}_{12}\tilde{b}_2 & \cdots & A_{11}^{k_2-1}b_1 + G\tilde{b}_2 & A_{11}g & \cdots & A_{11}^{k_1}g \\ \tilde{b}_2 & \tilde{A}_{22}\tilde{b}_2 & & \tilde{A}_{22}^{k_2-1}\tilde{b}_2 & 0 & & 0 \end{array}\right],$$

where G is a sum of products of A_{11}, $\tilde{A}_{12}$, and $\tilde{A}_{22}$, and $g = A_{11}^{k_2-1} + G\tilde{b}_2$. The positive entries of $\tilde{A}_{12}$ and $\tilde{A}_{22}$ may be chosen arbitrarily, but choose $\tilde{b}_2$ sufficiently small so that $b_1 \equiv (A_{11}^{k_2-1})^{-1}(u_{k_1} - G\tilde{b}_2)$ is sufficiently close to u_{k_1} and thus is positive. Then, since $g = u_{k_1}$, the last $k_1 \geqq 2$ columns of $C(A, b)$ are equal to $(u_{k_1}, 0)^T$, contradicting the controllability of (A, b). Thus k_1 cannot be $\geqq 2$, so A_{11} is of order 1 and (iii) is satisfied.

Case c. If $\tilde{A}_{22}$ is upper triangular with exactly one positive diagonal entry and A_{11} is a nonzero scalar or is an irreducible submatrix of order $\geqq 2$, then A does not satisfy (iii) and by Lemma 2, (A, b) is not controllable.

Thus, if **A** is an n-by-n nonnegative sign pattern and $(\mathbf{A}, \mathbf{B})$ is sign controllable, then (iii) is satisfied, completing the proof by induction. □

The only n-by-n nonnegative sign patterns **A** which satisfy the conditions of Theorem 1 (up to permutation similarity) are the following, where each $*$ (without restriction) may be 0 or +:

(i)

$$A = \left[\begin{array}{cc|ccccc} \square & \square & * & * & * & \cdots & * \\ 0 & 0 & + & * & * & \cdots & * \\ \hline & & 0 & + & * & \cdots & * \\ & & & 0 & + & \ddots & \vdots \\ 0 & & & & 0 & \ddots & * \\ & & & 0 & & \ddots & + \\ & & & & & & 0 \end{array}\right] \equiv \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ 0 & \mathbf{A}_{22} \end{bmatrix},$$

where

—**A** has order $n \geqq 2$;

—$\mathbf{A}_{11}$ must be present;

—$\mathbf{A}_{22}$ has order $n - 2 \geqq 0$ (and is void if $n = 2$);

—the unspecified entries denoted by □ must be such that

$$\operatorname{rank} \begin{bmatrix} \square & \square \\ 0 & 0 \end{bmatrix} = 1,$$

that is, this submatrix must be either

$$\begin{bmatrix} + & + \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & + \\ 0 & 0 \end{bmatrix}, \quad \text{or} \quad \begin{bmatrix} + & 0 \\ 0 & 0 \end{bmatrix}.$$

(ii)

$$A=\left[\begin{array}{ccccc|cc|ccccc}
0 & + & * & \cdots & * & * & * & * & * & * & \cdots & * \\
 & 0 & + & \ddots & \vdots & \vdots & \vdots & & & & & \\
 & & 0 & \ddots & * & * & * & \vdots & \vdots & \vdots & & \vdots \\
 & 0 & & \ddots & + & * & * & & & & & \\
 & & & & 0 & + & \square & * & * & * & \cdots & * \\
\hline
 & & 0 & & & + & \square & * & * & * & \cdots & * \\
 & & & & & 0 & 0 & + & * & * & \cdots & * \\
\hline
 & & & & & & & 0 & + & * & \cdots & * \\
 & & & & & & & & 0 & + & \ddots & \vdots \\
 & & 0 & & & 0 & & & & 0 & \ddots & * \\
 & & & & & & & & 0 & & \ddots & + \\
 & & & & & & & & & & & 0
\end{array}\right] \equiv \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix},$$

where

—**A** has order $n \geqq 3$;

—$\mathbf{A}_{11}$ has order $\geqq 1$ (and must be present);

—$\mathbf{A}_{22}$ must be present;

—$\mathbf{A}_{33}$ has order $n - 3 \geqq 0$ and may be void;

—the unspecified entries denoted by □ must be such that

$$\operatorname{rank}\begin{bmatrix} + & \square \\ + & \square \end{bmatrix} = 2,$$

that is, this submatrix must be either

$$\begin{bmatrix} + & + \\ + & 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} + & 0 \\ + & + \end{bmatrix}.$$

It is straightforward to test whether or not an n-by-n nonnegative sign pattern satisfies the conditions of Theorem 1. A pattern **A** satisfies the conditions if and only if it has precisely one zero row, which may be permuted (by permutation similarity) to row n. Then, consider the leading principal submatrix of order $k = n - 1$ of this permuted pattern; **A** satisfies the conditions of the theorem if and only if either

(a) this principal submatrix has precisely one zero row; or

(b) this principal submatrix has precisely one row (say row i) with a + in the diagonal position. Furthermore, if $k > 1$, all other entries of row i in this principal submatrix must be 0, there must be precisely one other row (say, row j) in this $k \times k$ principal submatrix which is identical to row i, and either one of the $(i, k + 1)$ and $(j, k + 1)$ entries of the permuted pattern **A** must be + and the other one must be 0.

If (a) is true, then permute (by permutation similarity) the zero row to row $n - 1$. Otherwise, permute the row i identified in (b) to row $n - 1$, by permutation similarity.

Now, similarly, continue consideration, in turn, of the leading principal submatrices of orders $k = n - 2, n - 3, \ldots, 1$ of the permuted pattern. Whenever possible, a zero row is permuted to row k of the array. If a + is ever permuted to the (k, k) entry of the array, then in the consideration of all remaining principal submatrices, a zero row must be able to be permuted to the last row of the principal submatrix (otherwise the pattern **A** does not satisfy the conditions of Theorem 1).

The above procedure for testing whether or not an arbitrary nonnegative sign pattern satisfies the conditions of the theorem is deterministic, and the permutation matrix P in the theorem is uniquely determined.

4. Other cases and open problems. In this paper we have considered only a special case of the sign controllability problem. Obvious extensions are the cases when $\mathbf{B}$ has zero entries (and $\mathbf{A}$ is nonnegative); when both $\mathbf{A}$ and $\mathbf{B}$ are sign patterns with entries in $\{+, -, 0\}$; and when $\mathbf{B}$ is an n-by-m sign pattern.

For the case when $\mathbf{A}$ and $\mathbf{B}$ are nonnegative and the vector pattern $\mathbf{B}$ has precisely one positive entry, we have the following sufficient condition for sign controllability.

THEOREM 2. *Let $n \geqq 2$, let $\mathbf{A}$ denote an n-by-n nonnegative sign pattern and let $\mathbf{B}$ denote an n-by-*1 *nonnegative sign pattern with precisely one positive entry. If there exists a permutation matrix P such that $P\mathbf{A}P^T$ is an unreduced upper Hessenberg pattern, i.e.,*

$$P\mathbf{A}P^T = \begin{bmatrix} * & * & \cdots & * & * \\ + & * & \cdots & * & * \\ & + & \ddots & \vdots & \vdots \\ 0 & & \ddots & * & * \\ & & & + & * \end{bmatrix}$$

(*where each* $*$ (*without restriction*) *may be* $+$ *or* 0), *and if* $P\mathbf{B} = [+, 0, \ldots, 0]^T$, *then* $(\mathbf{A}, \mathbf{B})$ *is sign controllable.*

Proof. If $A \in \mathbf{A}$ and $b \in \mathbf{B}$, then the controllability matrix $C(PAP^T, Pb)$ has the form

$$\begin{bmatrix} + & * & \cdots & * \\ & + & \ddots & \vdots \\ & & \ddots & * \\ 0 & & & + \end{bmatrix},$$

which clearly has full rank. □

The converse of this result, however, is not true, as the following example illustrates.

Example 2. Let

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & + \\ + & 0 & + \\ + & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} + \\ 0 \\ 0 \end{bmatrix}.$$

If $A \in \mathbf{A}$ and $b \in \mathbf{B}$, then $C(A, b)$ has the form

$$\begin{bmatrix} + & 0 & + \\ 0 & + & + \\ 0 & + & 0 \end{bmatrix},$$

which has negative determinant, implying that $(\mathbf{A}, \mathbf{B})$ is sign controllable. However, $\mathbf{A}$ is not permutation similar to an upper Hessenberg matrix (with a permutation that leaves $\mathbf{B}$ invariant).

Another obvious generalization of the sign controllability problem is to multi-input systems, i.e., systems in which the vector b is replaced by an n-by-m matrix B. Matrices A and B of orders n-by-n and n-by-m, respectively, are controllable if and only if the n-by-mn matrix $[B, AB, \ldots, A^{n-1}B]$ has rank n. Sign controllability is defined analogously to the case when B is a vector. The following is a direct consequence of Theorem 1.

THEOREM 3. *For $m \geqq 2$ and $n \geqq 2$, let $\mathbf{A}$ denote an n-by-n nonnegative sign pattern and let $\mathbf{B}$ denote an n-by-m nonnegative sign pattern. If $\mathbf{A}$ satisfies conditions* (i)–(iii) *of Theorem* 1 *and $\mathbf{B}$ has at least one nonzero entry in each of its rows, then* $(\mathbf{A}, \mathbf{B})$ *is sign controllable.*

Proof. Let $A \in \mathbf{A}$ and $B \in \mathbf{B}$. Since $b = Bu_m > 0$, by Theorem 1, $C(A, b)$ is controllable and thus has rank n. However, since

$$C(A, b) = C(A, B) \begin{bmatrix} u_m & 0 & \cdots & 0 \\ 0 & u_m & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & u_m \end{bmatrix},$$

$C(A, B)$ also has rank n. Thus $(\mathbf{A}, \mathbf{B})$ is sign controllable. □

The converse of Theorem 3 is also not true, as the following example illustrates.

Example 3. Let

$$\mathbf{A} = \begin{bmatrix} + & + \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ + & + \end{bmatrix}.$$

For all $A \in \mathbf{A}$ and $B \in \mathbf{B}$, the matrix $C(A, B)$ has the form

$$\begin{bmatrix} 0 & 0 & + & + \\ + & + & 0 & 0 \end{bmatrix},$$

which has rank 2.

Characterization of sign controllability, apart from the cases considered in Theorems 1, 2, and 3, remains an open problem.

Note added in proof. After this paper was accepted for publication, we discovered that the term "sign controllable" has been defined differently by Faibusovich (*Algebraic Riccati equation and symplectic algebra*, Internat. J. Control, 43 (1986), pp. 781–792). See also Scherer (*The solution set of the algebraic Riccati equation and the algebraic Riccati inequality*, Linear Algebra Appl., 153 (1991), pp. 99–122).

REFERENCES

[1] L. BASSETT, J. MAYBEE, AND J. QUIRK, *Qualitative economics and the scope of the correspondence principle*, Econometrica, 36 (1968), pp. 544–563.

[2] J. L. CASTI, *Linear Dynamical Systems*, Academic Press, Orlando, FL, 1987.

[3] C. R. JOHNSON, *Combinatorial matrix analysis: An overview*, Linear Algebra Appl., 107 (1988), pp. 3–15.

[4] V. KLEE, R. LADNER, AND R. MANBER, *Sign solvability revisited*, Linear Algebra Appl., 59 (1984), pp. 131–158.

[5] V. MEHRMANN AND G. M. KRAUSE, *Linear systems which leave controllable multi-input descriptor systems controllable*, Linear Algebra Appl., 120 (1989), pp. 47–64.

[6] K. MUROTA, *Systems Analysis by Graphs and Matroids*, Springer-Verlag, Berlin, 1987.

[7] V. M. POPOV, *Hyperstability of Control Systems*, Springer-Verlag, Berlin, 1973.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 408–419, April 1993

007

A NEW CRITERION TO GUARANTEE THE FEASIBILITY OF THE INTERVAL GAUSSIAN ALGORITHM*

A. FROMMER† AND G. MAYER‡

Abstract. The main subject of this paper is the interval Gaussian algorithm, which produces an interval vector $[x]^G$ analogously to its well-known counterpart in classical numerical analysis. Criteria are derived to guarantee the existence of $[x]^G$ if the $n \times n$ interval matrix $[A]$ is degenerate to a point matrix or if its comparison matrix $\langle [A] \rangle$ is irreducible and diagonally dominant. While in the first case all classical criteria of feasibility apply, the second case yields to a criterion which seems to be new. A way to construct matrices such that $[x]^G$ does not exist, although it does for each matrix $\tilde{A} \in [A]$, is also indicated.

Key words. linear interval equations, Gaussian algorithm, interval Gaussian algorithm, linear systems of equations, criteria of feasibility

AMS(MOS) subject classifications. 65F05, 65G10

1. Introduction. In many algorithms in interval analysis the interval Gaussian algorithm plays an important role (cf. [2]–[8], [15]–[17], [23]–[25]). Starting with an $n \times n$ interval matrix $[A]$ and a corresponding interval vector $[b]$, it produces an interval vector $[x]^G = \mathrm{IGA}([A], [b])$ which contains the solution set

$$S := \{\tilde{x} \in \mathbf{R}^n \mid \exists \tilde{A} \in [A], \tilde{b} \in [b] : \tilde{A}\tilde{x} = \tilde{b}\}.$$

The formulae to describe $[x]^G$ are given analogously to those of the well-known Gaussian elimination process in noninterval analysis, combined with a forward and backward substitution (cf. § 3). Unfortunately, as for real matrices $A \in \mathbf{R}^{n\times n}$, the interval Gaussian algorithm is not always feasible. If zero is contained in the first diagonal entry, e.g., the algorithm will necessarily break down. Pivoting, i.e., permuting rows and columns, may help, but, in contrast to nonsingular matrices $A \in \mathbf{R}^{n\times n}$, $[x]^G$ need not exist, even if pivoting is allowed and even if the Gaussian elimination process is feasible for each element $\tilde{A} \in [A]$ without pivoting (cf. [28] and [33]). While nonsingular real matrices $\tilde{A} \in \mathbf{R}^{n\times n}$ allow (theoretically) the elimination process without permutations if and only if all leading principal minors differ from zero (cf., e.g., § 4 or [34, Thm. 2.5, p. 120]), a similar criterion is still missing in interval analysis. Thus we look for classes of interval matrices that guarantee the existence of $[x]^G$ (cf. [1], [7], [9], [20], [30], [32], and [33]). An overview of such criteria can be found in [26].

In this paper we contribute further criteria for the feasibility of the interval Gaussian algorithm. We have two aspects in mind: First, we collect criteria for point matrices $[A]$, i.e., for *degenerate* interval matrices which consist of one element matrix only. Second, we choose some sort of limit case: It is known that $[x]^G$ exists for so-called H-matrices (cf. [1] or § 4). The set $\mathbf{H}$ of all such matrices is open in the Hausdorff-metric described in [7] (use Lemma 2.1 below for a proof). The boundary $\partial\mathbf{H}$ of $\mathbf{H}$ contains degenerate matrices $[A] = ([\tilde{a}_{ij}, \tilde{a}_{ij}])$ for which $\tilde{A} = (\tilde{a}_{ij})$ is irreducible and fulfills

$$|\tilde{a}_{ii}| = \sum_{\substack{j=1\\ j\neq i}}^{n} |\tilde{a}_{ij}|, \qquad i = 1, \ldots, n. \tag{1.1}$$

* Received by the editors March 4, 1991; accepted for publication (in revised form) August 30, 1991.

† Institut für Angewandte Mathematik, Universität Karlsruhe, Kaiserstraße 12, D-7500 Karlsruhe, Germany (ae09@dkauni2.bitnet).

‡ Fachbereich Mathematik, Bergische Universität, GH Wuppertal, Gaußstraße 20, D-5600 Wuppertal, Germany (frommer@math.uni-wuppertal.de).

As is seen by the example

$$\tilde{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

$[x]^G$ may no longer exist. We will discuss this case in greater detail, starting with necessary and sufficient conditions for the feasibility of the Gaussian algorithm. We extend the class of degenerate matrices which satisfy (1.1) onto a larger one contained in $\partial\mathbf{H}$. For this class, we will derive a sufficient condition for the existence of $[x]^G$. Again, we must answer negatively the question of whether the interval Gaussian algorithm is feasible for $[A]$ if its noninterval counterpart yields a vector $\tilde{x}^G$ for all elements $\tilde{A} \in [A]$. To this end we construct a counterexample giving some insight into the reasons that make the algorithm break down. The idea used for it can easily be generalized to obtain classes of matrices where $[x]^G$ does not exist.

We have arranged our paper as follows: Starting with some notation and basic facts of interval analysis in § 2, we continue by a short section presenting two possibilities for defining $[x]^G$. In § 4 we consider $[x]^G$ for degenerate matrices $[A]$; in § 5 we turn to the nondegenerate case using some generalization of condition (1.1). We illustrate our results by an example and discuss the counterexample mentioned above.

2. Preliminaries. By $\mathbf{R}^n$, $\mathbf{R}^{n\times n}$, $\mathbf{IR}$, $\mathbf{IR}^n$, and $\mathbf{IR}^{n\times n}$ we denote the set of real vectors with n components, the set of real $n \times n$ matrices, the set of intervals, the set of interval vectors with n components, and the set of $n \times n$ interval matrices, respectively. By "interval" we always mean a real compact interval. We write interval quantities in brackets with the exception of point quantities (i.e., degenerate interval quantities) which we identify with the element they contain. Examples are the null matrix 0 and the identity matrix I. We use the notation $[A] = [\underline{A}, \bar{A}] = ([a]_{ij}) = ([\underline{a}_{ij}, \bar{a}_{ij}]) \in \mathbf{IR}^{n\times n}$ simultaneously without further reference, and we proceed similarly for the elements of $\mathbf{R}^n$, $\mathbf{R}^{n\times n}$, $\mathbf{IR}$ and $\mathbf{IR}^n$.

By $A \geqq 0$ we denote a nonnegative $n \times n$ matrix, i.e., $a_{ij} \geqq 0$ for $i, j = 1, \ldots, n$. We call $x \in \mathbf{R}^n$ positive, writing $x > 0$ if $x_i > 0$, $i = 1, \ldots, n$.

We also mention the standard notation from interval analysis [7], [30]

$$\check{a} := \operatorname{mid}([a]) := \frac{\underline{a} + \bar{a}}{2} \quad \text{(midpoint)},$$

$$|[a]| := \max\{|\tilde{a}| \mid \tilde{a} \in [a]\} = \max\{|\underline{a}|, |\bar{a}|\} \quad \text{(absolute value)},$$

$$\langle [a] \rangle := \min\{|\tilde{a}| \mid \tilde{a} \in [a]\} = \begin{cases} \min\{|\underline{a}|, |\bar{a}|\} & \text{if } 0 \notin [a] \\ 0 & \text{otherwise} \end{cases} \quad \text{(minimal absolute value)}$$

for intervals $[a]$. For $[A] \in \mathbf{IR}^{n\times n}$ we obtain $\check{A}$, $|[A]| \in \mathbf{R}^{n\times n}$ by applying $\check{\ }$ and $|\cdot|$ entrywise, and we define the comparison matrix $\langle [A] \rangle = (c_{ij}) \in \mathbf{IR}^{n\times n}$ by setting

$$c_{ij} := \begin{cases} -|[a]_{ij}| & \text{if } i \neq j, \\ \langle [a_{ii}] \rangle & \text{if } i = j. \end{cases}$$

Since real quantities can be viewed as degenerate interval ones, $|\cdot|$ and $\langle \cdot \rangle$ can also be used for them. We call $[a] \in \mathbf{IR}$ symmetric (with respect to zero) if $[a] = -[a]$, i.e., if $[a] = [-|[a]|, |[a]|]$. Note that $\check{a} = 0$ if and only if $[a]$ is symmetric.

By $\mathbf{Z}^{n\times n}$ we denote the set of real $n \times n$ matrices with nonpositive off-diagonal entries; by $\det(A)$ we mean the determinant of a matrix $A \in \mathbf{R}^{n\times n}$.

As in [35, pp. 19, 20], we introduce the concept of a *directed graph* $G(A) := (\mathbf{X_A}, \mathbf{E_A})$ associated with a matrix $A \in \mathbf{R}^{n \times n}$. This graph consists of the set $\mathbf{X_A} := \{i \mid i = 1, \dots, n\}$ of nodes i and of the set $\mathbf{E_A} := \{(i, j) \mid a_{ij} \neq 0\}$ of edges (i, j), which are ordered pairs of nodes. A sequence of edges is termed a *path* of length r if it has the form $\{(i_l, i_{l+1})\}_{l=0}^{r-1}$. We will write $i_0 \to i_1 \to \cdots \to i_{r-1} \to i_r$ for this path. If there is a path $i =: i_0 \to i_1 \to \cdots \to i_r := j$ we say that i is *connected* to j. The matrix A is said to be *irreducible* if and only if any two nodes $i, j \in \mathbf{X_A}$ are connected. Note that A is irreducible if and only if $\langle A \rangle$ has this property.

In § 4 we will consider several classes of matrices $A \in \mathbf{R}^{n \times n}$, for which we are going to recall the definitions (cf. [11], [19], and [35]):

A is an *M-matrix* if A is nonsingular, $A^{-1} \geqq 0$, and $A \in \mathbf{Z}^{n \times n}$.

A is an *H-matrix* if $\langle A \rangle$ is an M-matrix.

A is *diagonally dominant* if for $i = 1, \dots, n$,

$$|a_{ii}| \geqq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \tag{2.1}$$

i.e., $\langle A \rangle e \geqq 0$, where here and in the sequel the vector e is defined by $e := (1, \dots, 1)^T \in \mathbf{R}^n$.

A is *strictly diagonally dominant* if (2.1) holds with strict inequality for all i.

A is *irreducibly diagonally dominant* if it is irreducible and diagonally dominant with strict inequality in (2.1) for at least one index i.

A is *totally nonnegative* if each minor of A is nonnegative.

An interval matrix $[A] \in \mathbf{IR}^{n \times n}$ is termed an *M-matrix* if each element $\tilde{A} \in [A]$ is an M-matrix. In the same way all other classes mentioned above can be extended to $\mathbf{IR}^{n \times n}$. There are several equivalent definitions for these classes of interval matrices. Thus it is easy to verify that

$[A]$ is an M-matrix if and only if $\underline{A}$ is an M-matrix and $\bar{a}_{ij} \leqq 0$ for $i \neq j$.

$[A]$ is an H-matrix if and only if $\langle [A] \rangle$ is an M-matrix.

To prove these two statements we can refer to a very useful criterion for M-matrices due to Fan [13].

LEMMA 2.1. *Let $A \in \mathbf{Z}^{n \times n}$. Then A is an M-matrix if and only if there exists a positive vector $x \in \mathbf{R}^n$ such that $Ax > 0$.*

We equip $\mathbf{IR}$, $\mathbf{IR}^n$, and $\mathbf{IR}^{n \times n}$ with the usual real interval arithmetic as described in [7], [10], [27], and [30], for example. We also take the "dos and don'ts" of this arithmetic for granted. We only mention the formulae (cf. [28])

$$\begin{cases} \langle [A] \pm [B] \rangle \geqq \langle [A] \rangle - |[B]|, \\ \left| \dfrac{1}{[a]} [A] \right| = \dfrac{1}{\langle [a] \rangle} |[A]| \quad \text{if } 0 \notin [a] \in \mathbf{IR}, \\ |[A] \pm [B]| \leqq |[A]| + |[B]|, \end{cases} \tag{2.2}$$

for interval matrices $[A]$, $[B]$. These formulae also hold, of course, for $n = 1$, i.e., for intervals. In this case they can be sharpened according to the following lemma, where here and in the sequel sign (a) is defined by

$$\operatorname{sign}(a) := \begin{cases} 1 & \text{if } a > 0, \\ 0 & \text{if } a = 0, \\ -1 & \text{if } a < 0 \end{cases}$$

for $a \in \mathbf{R}$.

LEMMA 2.2. *Let* $[a], [b], [c], [d] \in \mathbf{IR}$ *with* $0 \notin [d]$, *and define* $[a]' \in \mathbf{IR}$ *by*

$$[a]' := [a] - \frac{[b] \cdot [c]}{[d]}.$$

Then the following assertions hold.

(a) *The equation*

$$|[a]'| = |[a]| + \frac{|[b]| \cdot |[c]|}{\langle [d] \rangle} \tag{2.3}$$

is valid if and only if

$$\operatorname{sign}(\check{a}) \operatorname{sign}(\check{b}) \operatorname{sign}(\check{c}) \neq \operatorname{sign}(\check{d}). \tag{2.4}$$

In this case

$$\operatorname{sign}(\check{a}') = \operatorname{sign}(\check{a}), \quad \textit{provided } \check{a} \neq 0. \tag{2.5}$$

(b) *If we have*

$$\langle [a] \rangle \geqq \frac{|[b]| \cdot |[c]|}{\langle [d] \rangle}, \tag{2.6}$$

then

$$\langle [a]' \rangle = \langle [a] \rangle - \frac{|[b]| \cdot |[c]|}{\langle [d] \rangle} \tag{2.7}$$

is true if and only if

$$\operatorname{sign}(\check{a}) \operatorname{sign}(\check{b}) \operatorname{sign}(\check{c}) \neq -\operatorname{sign}(\check{d}). \tag{2.8}$$

In this case

$$\operatorname{sign}(\check{a}') = \operatorname{sign}(\check{a}), \quad \textit{provided } [a]' \neq 0. \tag{2.9}$$

Proof. (a) $0 \in \{\check{a}, \check{b}, \check{c}\}$ implies (2.3) and (2.4). Therefore, let $0 \notin \{\check{a}, \check{b}, \check{c}\}$. Then (2.3) is equivalent to

$$\operatorname{sign}(\check{a}) = -\operatorname{sign}\left(\operatorname{mid}\left(\frac{[b] \cdot [c]}{[d]}\right)\right),$$

hence, by elementary rules of interval arithmetic,

$$\operatorname{sign}(\check{a}) = -\frac{\operatorname{sign}(\check{b}) \operatorname{sign}(\check{c})}{\operatorname{sign}(\check{d})}.$$

This is, in turn, equivalent to (2.4). Equation (2.5) follows immediately from (2.3).

(b) $0 \in \{\check{b}, \check{c}\}$ implies $\check{a} = \check{a}'$ and (2.7)–(2.9). Therefore, let $0 \notin \{\check{b}, \check{c}\}$. Then $[b] \neq 0$, $[c] \neq 0$, and (2.6) guarantees $\langle [a] \rangle > 0$ which, in particular, means $\check{a} \neq 0$. Hence (2.7) is equivalent to

$$\operatorname{sign}(\check{a}) = \operatorname{sign}\left(\operatorname{mid}\left(\frac{[b] \cdot [c]}{[d]}\right)\right),$$

and this, in turn, is equivalent to (2.8). Equation (2.9) follows from (2.7) since $[a]' \neq 0$ implies that $\check{a}' \neq 0$ by (2.6) and (2.7). □

Note that sign $(\check{a})$ determines whether $|[a]| = \underline{a}$ or $|[a]| = \bar{a}$.

3. The algorithm. As in the noninterval case the interval vector $[x]^G$ resulting from the interval Gaussian algorithm (which we will also denote by IGA$([A], [b])$) can be described using the following matrices $[A]^{(k)} \in \mathbf{IR}^{n\times n}$ and vectors $[b]^{(k)} \in \mathbf{IR}^n$, $k = 1, \ldots, n$:

$$[A]^{(1)} := [A], \qquad [b]^{(1)} := [b];$$

for $k = 1, \ldots, n-1$,

$$(3.1)\quad \begin{aligned} [a]_{ij}^{(k+1)} &:= \begin{cases} [a]_{ij}^{(k)} & i = 1,\ldots,k;\ \ j = 1,\ldots,n, \\ [a]_{ij}^{(k)} - \dfrac{[a]_{ik}^{(k)}\cdot[a]_{kj}^{(k)}}{[a]_{kk}^{(k)}} & i = k+1,\ldots,n;\ \ j = k+1,\ldots,n, \\ 0 & \text{otherwise}, \end{cases} \\ [b]_i^{(k+1)} &:= \begin{cases} [b]_i^{(k)} & i = 1,\ldots,k, \\ [b]_i^{(k)} - \dfrac{[a]_{ik}^{(k)}}{[a]_{kk}^{(k)}}\cdot[b]_k^{(k)} & i = k+1,\ldots,n. \end{cases} \end{aligned}$$

Then

$$[x]_i^G := \left([b]_i^{(n)} - \sum_{j=i+1}^{n} [a]_{ij}^{(n)}[x]_j^G\right) \Big/ [a]_{ii}^{(n)}, \qquad i = n(-1)1,$$

where $\sum_{j=n+1}^{n} \cdots$ is defined as zero. We emphasize that in the whole paper, $[x]^G$ is always defined according to the expressions above, i.e., without permuting rows or columns. Pivoting can be found in [22], but satisfactory results are still missing. The construction of $[A]^{(n)}$, $[b]^{(n)}$, and $[x]^G$ corresponds to the triangular decomposition, the forward substitution, and the backward substitution, respectively, in the noninterval case. The way of constructing the matrices $[A]^{(k)}$ can also be described using the Schur complement. Since our proof in § 5 is based on the Schur complement we review it shortly in the following definition.

DEFINITION 3.1. (Cf. [28].) Let

$$[A] := \begin{pmatrix} [a]_{11} & [c]^T \\ [d] & [A]' \end{pmatrix} \in \mathbf{IR}^{n\times n}, \qquad n > 1,$$

with $[c], [d] \in \mathbf{IR}^{n-1}$ and $[A]' \in \mathbf{IR}^{(n-1)\times(n-1)}$.

(a) The $(n-1)\times(n-1)$ interval matrix

$$(3.2)\qquad \Sigma_{[A]} := [A]' - \frac{1}{[a]_{11}}[d][c]^T$$

is termed *Schur complement* (of the (1, 1) entry), provided that $0 \notin [a]_{11}$. $\Sigma_{[A]}$ is not defined if $0 \in [a]_{11}$.

(b) $[A]$ has the *triangular decomposition* $([L], [U])$ if either $n = 1$ and $[L] = 1$, $[U] = [A] \not\ni 0$, or if $n > 1$ and

$$[L] := \begin{pmatrix} 1 & 0 \\ \dfrac{1}{[a]_{11}}[d] & [L]' \end{pmatrix}, \qquad [U] := \begin{pmatrix} [a]_{11} & [c]^T \\ 0 & [U]' \end{pmatrix},$$

where $0 \notin [a]_{11}$ and $([L]', [U]')$ is the triangular decomposition of $\Sigma_{[A]}$.

Note that, as opposed to the noninterval case, $[L]\cdot[U] = [A]$ is not always true. It is obvious that the following assertions are equivalent.

(a) The vector IGA $([A], [b])$ exists for all $[b] \in \mathbf{IR}^n$.

(b) The matrices $[A]^{(k)}$, $k = 1, \ldots, n$, exist with $0 \notin [a]_{nn}^{(n)}$.

(c) The triangular decomposition $([L], [U])$ of $[A]$ exists.

(d) $0 \notin [a]_{kk}^{(k)}$, $k = 1, \ldots, n$.

It is also important to note that the feasibility of the interval Gaussian algorithm (3.1) (but not the result) only depends on the matrix $[A]$ and not on the vector $[b]$.

Hereafter, we will sometimes use the matrix

$$L_{[A]} := \begin{pmatrix} 1 & & & \\ -\dfrac{[a]_{21}}{[a]_{11}} & 1 & & 0 \\ \vdots & & \ddots & \\ -\dfrac{[a]_{n1}}{[a]_{11}} & 0 & & 1 \end{pmatrix} \in \mathbf{IR}^{n \times n}, \tag{3.3}$$

which is defined if $0 \notin [a]_{11}$. Obviously,

$$[a]_{ij}^{(2)} = (L_{[A]} \cdot [A])_{ij}, \qquad i, j = 2, \ldots, n.$$

To shorten some proofs in the sections to follow we state two lemmata of purely auxiliary character.

LEMMA 3.2. *Let* $[A] \in \mathbf{IR}^{n\times n}$, $n \geqq 2$, $0 \notin [a]_{11}$, $0 < x \in \mathbf{R}^n$. *Then the following assertions hold*:

(a) $\langle \Sigma_{[A]} \rangle \geqq \Sigma_{\langle [A] \rangle}$ $(\in \mathbf{Z}^{(n-1)\times(n-1)}$, *if* $n \geqq 3)$.

(b) $\langle [A]^{(2)} \rangle \geqq \langle [A] \rangle^{(2)}$.

(c) $\langle [A] \rangle x \left\{\begin{matrix} \geqq \\ > \\ = \end{matrix}\right\} 0$ *implies* $\langle [A]^{(2)} \rangle x \geqq \langle [A] \rangle^{(2)} x \left\{\begin{matrix} \geqq \\ > \\ = \end{matrix}\right\} 0.$

Proof. (a) Using (3.2) we obtain by (2.2)

$$\langle \Sigma_{[A]} \rangle = \left\langle [A]' - \frac{1}{[a]_{11}} [d][c]^T \right\rangle \geqq \langle [A]' \rangle - \frac{1}{\langle [a]_{11} \rangle} |[d]|\,|[c]^T|$$
$$= \Sigma_{\langle [A] \rangle} \quad \text{(cf. also [28, Prop. 6i])}.$$

Assertion (b) follows directly from (a) and

$$\langle [A]^{(2)} \rangle = \begin{pmatrix} \langle [a]_{11} \rangle & -|[a]_{12}| & \cdots & -|[a]_{1n}| \\ 0 & & & \\ \vdots & & \langle \Sigma_{[A]} \rangle & \\ 0 & & & \end{pmatrix}.$$

(c) Since $\langle [A] \rangle^{(2)} = L_{\langle [A] \rangle} \cdot \langle [A] \rangle$ with $L_{\langle [A] \rangle}$ defined for $\langle [A] \rangle$ analogously to (3.3), we obtain, together with (b),

$$\langle [A]^{(2)} \rangle x \geqq L_{\langle [A] \rangle}(\langle [A] \rangle x) \geqq I \cdot \langle [A] \rangle x,$$

thus establishing (c). □

LEMMA 3.3. *Let* $[A] \in \mathbf{IR}^{n \times n}$ *satisfy the following properties*:

$$(3.4) \qquad \langle [A] \rangle \text{ is irreducible,}$$

$$(3.5) \qquad \langle [A] \rangle e \geqq 0.$$

Then we get

(a) $\langle [a]_{ii} \rangle > 0$, $i = 1, \ldots, n$.

(b) $\Sigma_{\langle [A] \rangle}$ *is irreducible, provided* $n \geqq 3$.

Proof. (a) If $n = 1$ the assertion follows by the definition of the irreducibility. Therefore, let $n > 1$. By (3.4), for any $i \in \{1, \ldots, n\}$ there exists an index $j = j(i) \neq i$ such that $[a]_{ij} \neq 0$. Hence (3.5) implies that

$$0 \leqq \langle [a]_{ii} \rangle - |[a]_{ij}| < \langle [a]_{ii} \rangle .$$

(b) Choose any two indices $i, j \in \{2, \ldots, n\}$, assuming $i \neq j$. (Note $n \geqq 3$ by assumption!) Then

$$(3.6) \qquad (\langle [A] \rangle^{(2)})_{ij} = -|[a]_{ij}| - \frac{|[a]_{i1}|\,|[a]_{1j}|}{\langle [a]_{11} \rangle},$$

and thus

$$(3.7) \qquad (\langle [A] \rangle^{(2)})_{ij} < 0 \Leftrightarrow [a]_{ij} \neq 0 \quad \text{or} \quad ([a]_{i1} \neq 0 \text{ and } [a]_{1j} \neq 0).$$

Choose a shortest path in the directed graph $G(\langle [A] \rangle)$ which connects the node i to the node j, say,

$$(3.8) \qquad i =: i_1 \rightarrow i_2 \rightarrow \cdots \rightarrow i_{s-1} \rightarrow i_s := j.$$

Case 1. $i_l \neq 1$ for all $l \in \{1, \ldots, s\}$. Then by (3.7) $(\langle [A] \rangle^{(2)})_{i_l i_{l+1}} \neq 0$, $l = 1, \ldots, s-1$, hence (3.8) is a path in $G(\langle [A] \rangle^{(2)})$ which connects the node i to the node j.

Case 2. $i_l = 1$ for some index $l \in \{1, \ldots, s\}$. Since the path of (3.8) is of minimal length, l is the only index for which $i_l = 1$. Then $l \notin \{1, s\}$ and $[a]_{i_{l-1}1} \neq 0$, $[a]_{1i_{l+1}} \neq 0$, hence

$$(\langle [A] \rangle^{(2)})_{i_{l-1}, i_{l+1}} < 0$$

by (3.7). Therefore, the path

$$i = i_1 \rightarrow i_2 \rightarrow \cdots \rightarrow i_{l-1} \rightarrow i_{l+1} \rightarrow \cdots \rightarrow i_s = j$$

connects the node i to the node j in $G(\langle [A] \rangle^{(2)})$.

Thus we have shown that for any nodes $i, j \in \{2, \ldots, n\}$, $i \neq j$, there is a path in $G(\langle [A] \rangle^{(2)})$ connecting i to j without containing the node 1. By concatenating the paths from i to any node $l \in \{2, \ldots, n\} \backslash \{i\}$ and from l to i, this assertion holds also for the case $i = j$. Hence, any two nodes are connected in $G(\Sigma_{\langle [A] \rangle})$ and thus $\Sigma_{\langle [A] \rangle}$ is irreducible. □

4. The degenerate case. In this section we consider the feasibility of the interval Gaussian algorithm if $[A]$ is a degenerate interval matrix. As we have already pointed out, this feasibility does not depend on $[b]$. Therefore, it is clear that we can adopt all criteria that are known in the literature for the Gaussian elimination process in the noninterval case. We start with the following well-known necessary and sufficient conditions which can be found in [34].

THEOREM 4.1. *Let* $A \in \mathbf{R}^{n \times n}$. *Then* IGA $(A, [b])$ *exists for any vector* $[b] \in \mathbf{IR}^n$ *if and only if* $\det(A_k) \neq 0$, $k = 1, \ldots, n$, *where* A_k *is the leading principal submatrix of* A *of order* k.

Based on Theorem 4.1, we can immediately find classes of matrices for which $[x]^G$ exists. We list some of them in our next theorem.

THEOREM 4.2. *Let* $A \in \mathbf{R}^{n \times n}$, $[b] \in \mathbf{IR}^n$. *Then* IGA $(A, [b])$ *exists in each of the following cases*:

(a) *A is symmetric and positive definite.*

(b) *A is an H-matrix.*

(c) *A is nonsingular and diagonally dominant.*

(d) *A is nonsingular and totally nonnegative.*

We omit the proof, referring the reader to [21] for (a), to [18] or [14] for (b) and (c), and to [12] for (d).

In view of § 5 we consider irreducible matrices A which are diagonally dominant. For such matrices we recall a necessary and sufficient criterion for A to be singular, so that Theorem 4.2(c) can be applied in all other cases.

THEOREM 4.3. *Let* $A \in \mathbf{R}^{n \times n}$, $[b] \in \mathbf{IR}^n$, *A being irreducible and diagonally dominant. Then*

$$a_{ii} \neq 0, \qquad i = 1, \ldots, n. \tag{4.1}$$

Furthermore, A is singular if and only if $\langle A \rangle e = 0$ *and*

$$A = D_\sigma D_A \langle A \rangle D_\sigma, \tag{4.2}$$

where $D_\sigma \in \mathbf{R}^{n \times n}$, $|D_\sigma| = I$, *and* $D_A = \operatorname{diag}(\operatorname{sign}(a_{11}), \ldots, \operatorname{sign}(a_{nn}))$. *Therefore,* IGA $(A, [b])$ *exists if and only if A is not representable in the form* (4.2).

A generalization of this theorem can be found in [31], as was pointed out to us by one of the referees; use "$A \in \mathbf{R}^{n \times n}$" in the proof of Satz V in [31] to show $D_\sigma \in \mathbf{R}^{n \times n}$.

5. The general case. In this section we drop the restriction of $[A]$ being degenerate. As mentioned in § 1, $[x]^G$ does not necessarily exist if the Gaussian algorithm is feasible for each element matrix $\tilde{A} \in [A]$. We might conjecture that at least Theorem 4.2 is generalizable to nondegenerate interval matrices. Reichmann's example in [33] shows that this is not the case for symmetric positive definite matrices $[A]$. (These are defined by $[A] = [A]^T$ and by $\tilde{A}$ being positive definite if $\tilde{A} \in [A]$ is symmetric.) This is astonishing, since Wilkinson showed in [36] that $\Sigma_{\tilde{A}}$ is again symmetric positive definite, provided $\tilde{\mathbf{A}}$ has this property. Garloff gives in [20] a counterexample which shows that Theorem 4.2(d) does not generally hold for nondegenerate interval matrices. We will illustrate by Example 5.5 below that Theorem 4.2(c) can also fail when $\underline{A} \neq \bar{A}$. For H-matrices, Alefeld showed in [1] the existence of $[x]^G$. We recall this important phenomenon as Theorem 5.1.

THEOREM 5.1. *Let* $[A] \in \mathbf{IR}^{n \times n}$ *be an H-matrix and let* $[b] \in \mathbf{IR}^n$. *Then* $[x]^G$ *exists.*

Using [14, Satz B.2.5] we can easily show that the matrices in the following corollary are special H-matrices.

COROLLARY 5.2. *Let* $[A] \in \mathbf{IR}^{n \times n}$ *and* $[b] \in \mathbf{IR}^n$. *In each of the following cases,* $[A]$ *is an H-matrix and thus* $[x]^G$ *exists*:

(a) $[A]$ *is an M-matrix.*

(b) $\langle [A] \rangle$ *is strictly diagonally dominant.*

(c) $\langle [A] \rangle$ *is irreducibly diagonally dominant.*

(d) $\langle [A] \rangle$ *is nonsingular and diagonally dominant.*

As motivated in § 1 we now consider the following weakening of Corollary 5.2(c) and, with restrictions, of Corollary 5.2(d).

THEOREM 5.3. *Let* $[A] \in \mathbf{IR}^{n\times n}$, $n \geqq 2$, *and* $[b] \in \mathbf{IR}^n$. *Assume that*

$$\langle [A] \rangle \text{ is irreducible} \tag{5.1}$$

and

$$\langle [A] \rangle e \geqq 0. \tag{5.2}$$

If

$$\operatorname{sign}(\check{a}_{ij}) \cdot \operatorname{sign}(\check{a}_{ik}) \cdot \operatorname{sign}(\check{a}_{kj}) = \begin{cases} \operatorname{sign}(\check{a}_{kk}) & \text{if } i \neq j, \\ -\operatorname{sign}(\check{a}_{kk}) & \text{if } i = j \end{cases} \tag{5.3}$$

for at least one triple $i, j, k \in \{1, \ldots, n\}$ *satisfying* $k < i, j$, *then* $\Sigma_{[A]}$ *exists. Moreover,* $\Sigma_{[A]}$ *is an H-matrix or it fulfills again the conditions* (5.1)–(5.3). *In particular, these properties imply the existence of* $[x]^G$.

Before proving Theorem 5.3 we give a simple example illustrating its contents.

Example 5.4. Define $[A]_\alpha$ by

$$[A]_\alpha := \begin{pmatrix} 2 & 1 & -1 \\ [\alpha, 1] & 2 & -1 \\ [-1, 1] & -1 & 2 \end{pmatrix}, \quad \text{where } \alpha \in (-1, 1].$$

$[A]_\alpha$ is not an H-matrix since $\langle [A]_\alpha \rangle$ is singular ($\langle [A]_\alpha \rangle e = 0$). The sign-matrix

$$S_{[A]} := (\operatorname{sign}(\check{a}_{ij})) \in \mathbf{R}^{3\times 3}$$

is given by

$$S_{[A]} = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

The product of the entries s_{13}, s_{21}, and s_{23} is equal to the entry s_{11}. Thus (5.3) is fulfilled for $(i, j, k) = (2, 3, 1)$, and by Theorem 5.3, IGA $([A]_\alpha, [b])$ exists for any $[b] \in \mathbf{IR}^n$. Note that $\alpha \neq -1$ is necessary to guarantee the feasibility of the interval Gaussian algorithm, as will be shown by Example 5.5 below.

Proof of Theorem 5.3. We may assume that the only degenerate entries of $[A]$ are zeros. (Otherwise, we can slightly blow up the nonzero degenerate entries to obtain a matrix $[C]$ satisfying

$$[A] \subseteq [C], \quad \langle [A] \rangle = \langle [C] \rangle, \quad \operatorname{sign}(\check{a}_{ij}) = \operatorname{sign}(\check{c}_{ij}), \quad i, j = 1, \ldots, n.$$

In this way the assumptions of Theorem 5.3 hold also for $[C]$, and we prove Theorem 5.3 for $[C]$ instead of $[A]$. Clearly, the interval Gaussian algorithm is feasible for $[A]$ if it is for $[C]$.) So, from now on we always assume that $[A]$ has only zero or nondegenerate entries. Thus for $1 < i, j$ we have

$$[a]_{ij}^{(2)} = 0 \Leftrightarrow [a]_{ij} = 0 \quad \text{and} \quad ([a]_{i2} = 0 \text{ or } [a]_{2j} = 0)$$

and

$$[a]_{ij} \neq 0 \Rightarrow [a]_{ij}^{(2)} \neq 0.$$

In particular, $\langle \Sigma_{[A]} \rangle$ is irreducible, if $\Sigma_{\langle [A] \rangle}$ has this property. In addition, the only degenerate entries of $\Sigma_{[A]}$ are again zeros.

If $(\langle [A] \rangle e)_i > 0$ for at least one index $i \in \{1, \ldots, n\}$, then $[x]^G$ exists by Corollary 5.2(c), and $[A]$ is an H-matrix. Thus we assume, without loss of generality, that $\langle [A] \rangle e = 0$ holds.

Let $n = 2$. Then in (5.3) $k = 1$, $i = j = 2$ necessarily holds. By Lemma 3.3(a),

$$\Sigma_{[A]} = \left([a]_{22} - \frac{[a]_{21}[a]_{12}}{[a]_{11}} \right)$$

exists, and $0 \notin [a]_{22}$. Together with (5.3) this implies

$$0 \neq \operatorname{sign}(\check{a}_{22}) = -\frac{\operatorname{sign}(\check{a}_{21})\cdot\operatorname{sign}(\check{a}_{12})}{\operatorname{sign}(\check{a}_{11})} = -\operatorname{sign}\left(\operatorname{mid}\left(\frac{[a]_{21}[a]_{12}}{[a]_{11}}\right)\right).$$

Hence, by Lemma 2.2(b),

$$\langle [a]_{22}^{(2)} \rangle > \langle [a]_{22} \rangle - \frac{|[a]_{21}|\,|[a]_{12}|}{\langle [a]_{11} \rangle},$$

the right-hand side being equal to zero by Lemma 3.2(c) (with $x = e$). Thus $\Sigma_{[A]} \in \mathbf{IR}^{1\times 1}$ is an H-matrix and $[x]^G$ exists.

Now let $n \geqq 3$. By Lemma 3.3(a), $\Sigma_{[A]}$ exists, and by Lemma 3.2(a) and (c) we get

$$\langle \Sigma_{[A]} \rangle e' \geqq \Sigma_{\langle [A] \rangle} e' = 0,$$

where $e' = (1, \ldots, 1)^T \in \mathbf{R}^{n-1}$. Since $\Sigma_{\langle [A] \rangle}$ is irreducible by Lemma 3.3(b), so is $\langle \Sigma_{[A]} \rangle$.

Case 1. There is an index $i_0 \in \{1, \ldots, n-1\}$ such that

$$(5.4) \qquad (\langle \Sigma_{[A]} \rangle e')_{i_0} > 0.$$

Then $\langle \Sigma_{[A]} \rangle$ is irreducibly diagonally dominant, thus $\Sigma_{[A]}$ is an H-matrix.

Case 2.

$$(5.5) \qquad \langle \Sigma_{[A]} \rangle e' = 0.$$

By Lemma 3.2(a), (c) we get

$$0 = \langle \Sigma_{[A]} \rangle e' \geqq \Sigma_{\langle [A] \rangle} e' = 0,$$

hence

$$(5.6) \qquad \langle \Sigma_{[A]} \rangle = \Sigma_{\langle [A] \rangle}.$$

Assumption (5.3) and Lemma 3.3(a) imply

$$(5.7) \qquad 0 \notin \{\check{a}_{ij}, \check{a}_{ik}, \check{a}_{kj}, \check{a}_{kk}\},$$

and Lemma 2.2 and equations (5.3) and (5.6) yield $k > 1$. From Lemma 2.2 and equations (5.6) and (5.7) we also get

$$\operatorname{sign}(\check{a}_{ij}) = \operatorname{sign}(\check{a}_{ij}^{(2)}),$$
$$\operatorname{sign}(\check{a}_{ik}) = \operatorname{sign}(\check{a}_{ik}^{(2)}),$$
$$\operatorname{sign}(\check{a}_{kj}) = \operatorname{sign}(\check{a}_{kj}^{(2)}),$$
$$\operatorname{sign}(\check{a}_{kk}) = \operatorname{sign}(\check{a}_{kk}^{(2)}).$$

Hence (5.3) holds for $\Sigma_{[A]}$; (5.1) was stated for this matrix already and (5.2) follows from (5.5) for it.

So far we have shown that $\Sigma_{[A]}$ exists and is an H-matrix or satisfies the assumptions of Theorem 5.3 in the place of $[A]$. In the first case, $[x]^G$ exists by Theorem 5.1. In the second case, we can again apply the argumentation above. Since we have already seen the theorem to be true for $n = 2$, this concludes our proof by an induction of at most $n - 1$ steps. □

The conditions (5.1) and (5.2) of Theorem 5.3 cause $[A]$ either to be an H-matrix or to satisfy $\langle [A] \rangle e = 0$. In this latter case the existence of $[x]^G$ implies that $[A]$ does not contain any matrix $\tilde{A}$ which fulfills

$$\tilde{A} = D_\sigma D_{\tilde{A}} \langle [A] \rangle D_\sigma,$$

where D_σ, $D_{\tilde{A}}$ are defined as in (4.2) (cf. Theorem 4.3). However, this is not sufficient to guarantee the existence of $[x]^G$, as can be seen by the following example.

Example 5.5. Consider the matrix $[A]_{-1}$ from Example 5.4, i.e.,

$$[A] := [A]_{-1} = \begin{pmatrix} 2 & 1 & -1 \\ [-1, 1] & 2 & -1 \\ [-1, 1] & -1 & 2 \end{pmatrix}.$$

$\langle [A] \rangle$ is irreducible, $\langle [A] \rangle e = 0$, but condition (5.3) is not fulfilled. Since

$$\det \begin{pmatrix} 2 & 1 & -1 \\ \alpha & 2 & -1 \\ \beta & -1 & 2 \end{pmatrix} = 6 - \alpha + \beta \geqq 4 \neq 0 \qquad (\alpha, \beta \in [-1, 1]),$$

each element matrix $\tilde{A}$ of $[A]$ is nonsingular, hence none of them can be represented by (4.2). Theorem 4.2(c) guarantees the existence of IGA $(\tilde{A}, [b])$ for any $[b] \in \mathbf{IR}^n$. Nevertheless, IGA $([A], [b])$ does not exist since $[A]$ and

$$[B] := \begin{pmatrix} 2 & [-1, 1] & -1 \\ [-1, 1] & 2 & -1 \\ [-1, 1] & -1 & 2 \end{pmatrix}$$

produce the same Schur complement $\Sigma_{[A]} = \Sigma_{[B]}$. This is due to the equality

$$[a]_{21}[a]_{12} = [-1, 1] \cdot 1 = [-1, 1] \cdot [-1, 1] = [b]_{21}[b]_{12}.$$

The products occur when computing $[A]^{(2)}$ and $[B]^{(2)}$, respectively. Since $\langle [B] \rangle \in [B]$ and $\langle [B] \rangle e = 0$, the matrix $[B]$ contains a singular matrix, hence IGA $([B], [b])$, and therefore IGA $([A], [b])$, cannot exist.

The idea of a type of backward analysis used to construct a matrix $[B]$ with $\Sigma_{[B]} = \Sigma_{[A]}$ can certainly be generalized, and can be used to construct classes of interval matrices for which IGA $([A], [b])$ does not exist. We do not pursue this aspect further here.

Acknowledgment. The authors are very grateful to the referees for their valuable suggestions.

REFERENCES

[1] G. ALEFELD, *Über die Durchführbarkeit des Gaußchen Algorithmus bei Gleichungen mit Intervallen als Koeffizienten*, Comput. Suppl., 1 (1977), pp. 15–19.

[2] ———, *Intervallanalytische Methoden bei nichtlinearen Gleichungen*, in Jahrbuch Überblicke Mathematik, 1979, Bibliographisches Institut, Mannheim, 1979, pp. 63–78.

[3] ———, *On the convergence of some interval-arithmetic modifications of Newton's method*, SIAM J. Numer. Anal., 21 (1984), pp. 363–372.

[4] ———, *Berechenbare Fehlerschranken für ein Eigenpaar unter Einschluss von Rundungsfehlern bei Verwendung des genauen Skalarprodukts*, Z. Angew. Math. Mech., 67 (1987), pp. 145–152.

[5] ———, *Rigorous error bounds for singular values of a matrix using the precise scalar product*, in Com-

puterarithmetic, E. Kaucher, U. Kulisch, and Ch. Ullrich, eds., B. G. Teubner, Stuttgart, 1987, pp. 9-30.

[6] G. ALEFELD, *Error bounds for quadratic systems of nonlinear equations using the precise scalar product*, in Scientific Computation with Automatic Result Verification, U. Kulisch and H. J. Stetter, eds., Comput. Suppl., 6, Springer-Verlag, Vienna, 1988, pp. 59-68.

[7] G. ALEFELD AND J. HERZBERGER, *Introduction to Interval Computations*, Academic Press, New York, 1983.

[8] G. ALEFELD AND L. PLATZÖDER, *A quadratically convergent Krawczyk-like algorithm*, SIAM J. Numer. Anal., 20 (1983), pp. 210-219.

[9] W. BARTH AND E. NUDING, *Optimale Lösung von Intervallgleichungssystemen*, Computing, 12 (1974), pp. 117-125.

[10] H. BAUCH, K.-U. JAHN, D. OELSCHLÄGEL, H. SÜSSE, AND V. WIEBIGKE, *Intervallmathematik*, BSB B.G. Teubner Verlagsgesellschaft, Leipzig, 1987.

[11] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[12] C. W. CRYER, *The* LU-*factorization of totally positive matrices*, Linear Algebra Appl., 7 (1973), pp. 83-92.

[13] K. FAN, *Topological proof for certain theorems on matrices with nonnegative elements*, Monatsh. Math., 62 (1958), pp. 219-237.

[14] A. FROMMER, *Lösung linearer Gleichungssysteme auf Parallelrechnern*, Vieweg, Braunschweig, 1990.

[15] A. FROMMER AND G. MAYER, *Parallel interval multisplittings*, Numer. Math., 56 (1989), pp. 255-267.

[16] ———, *Safe bounds for the solution of nonlinear problems using a parallel multisplitting method*, Computing, 42 (1989), pp. 171-186.

[17] ———, *Efficient methods for enclosing solutions of systems of nonlinear equations*, Computing, 44 (1990), pp. 221-235.

[18] R. E. FUNDERLIC, M. NEUMANN, AND R. J. PLEMMONS, LU *decompositions of generalized diagonally dominant matrices*, Numer. Math., 40 (1982), pp. 57-69.

[19] F. S. GANTMACHER, *Matrizentheorie*, Springer-Verlag, Berlin, 1986.

[20] J. GARLOFF, *Totally nonnegative interval matrices*, in Interval Mathematics 1980, K. L. E. Nickel, ed., Academic Press, New York, 1980, pp. 317-327.

[21] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.

[22] M. HEBGEN, *Eine scaling-invariante Pivotsuche für Intervallmatrizen*, Computing, 12 (1974), pp. 99-106.

[23] G. MAYER, *Reguläre Zerlegungen und der Satz von Stein und Rosenberg für Intervallmatrizen*, Habilitationsschrift, Universität Karlsruhe, Karlsruhe, 1986.

[24] ———, *On the speed of convergence of some iterative processes*, in Colloquia Mathematica Societatis János Bolyai, 50, Numerical Methods, Miskolc, 1986, D. Greenspan and P. Rósza, eds., North-Holland, Amsterdam, 1988, pp. 207-228.

[25] ———, *On Newton-like methods to enclose solutions of nonlinear equations*, Appl. Math., 34 (1989), pp. 67-84.

[26] ———, *Old and new aspects of the interval Gaussian algorithm*, in Computer Arithmetic, Scientific Computation and Mathematical Modelling, E. Kaucher, S. M. Markov, and G. Mayer, eds., IMACS Annals on Computing and Applied Mathematics, Baltzer, Basel, 1991.

[27] R. E. MOORE, *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1966.

[28] A. NEUMAIER, *New techniques for the analysis of linear interval equations*, Linear Algebra Appl., 58 (1984), pp. 273-325.

[29] ———, *Further results on linear interval equations*, Linear Algebra Appl., 87 (1987), pp. 155-179.

[30] ———, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, 1990.

[31] A. OSTROWSKI, *Über die Determinanten mit überwiegender Hauptdiagonale*, Comment. Helv. Math., 10 (1937), pp. 69-96.

[32] K. REICHMANN, *Ein hinreichendes Kriterium für die Durchführbarkeit des Intervall-Gauss-Algorithmus bei Intervall-Hessenberg-Matrizen ohne Pivotsuche*, Z. Angew. Math. Mech., 59 (1975), pp. 373-379.

[33] ———, *Abbruch beim Intervall-Gauss-Algorithmus*, Computing, 22 (1979), pp. 355-361.

[34] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[35] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[36] J. H. WILKINSON, *Error analysis of direct methods for matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281-330.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 420–431, April 1993

008

EFFICIENT COMPUTATION OF THE SOLUTIONS TO MODIFIED LYAPUNOV EQUATIONS*

STEPHEN RICHTER†, LARRY D. DAVIS†, AND EMMANUEL G. COLLINS, JR.†

Abstract. This paper develops a solution method for modified Lyapunov equations in which the modification term $\mathscr{F}(Q)$ is a linear function of the solution Q. Equations of this form arise in robustness analysis and in homotopy algorithms developed for solving the nonstandard Riccati and Lyapunov equations that arise in robust reduced-order design. The method relies on decomposing $\mathscr{F}(Q)$ as $\mathscr{F}(Q) = \mathscr{G}(\phi(Q))$, where $\phi(Q)$ is an m-dimensional vector. It is shown that if m is small, the new solution procedure is much more efficient than are solution procedures based on a straightforward transformation of the modified Lyapunov equation to a linear vector equation in $n(n+1)/2$ unknowns. The results are extended to develop an efficient procedure for computing the solutions to an arbitrary number of coupled Lyapunov equations in which the coupling terms are linear operators.

Key words. Lyapunov equations, linear systems of equations

AMS(MOS) subject classifications. 15A06, 15A24, 65F05

Nomenclature. The following summarizes the nomenclature used in this paper.

$\mathbb{S}^{n\times n}$	the space of symmetric matrices in $\mathbb{R}^{n\times n}$.
$X_1^r\mathbb{S}^{n\times n}$	$\mathbb{S}^{n\times n}\times\cdots\times\mathbb{S}^{n\times n}$ ($r-1$ cross products).
s_{ij} or $S_{i,j}$	the (i,j) element of the matrix S.
I_m	the $m\times m$ identity matrix.
$0_{m\times n}$	the $m\times n$ zero matrix.
$e_m^{(j)}$	the mth-order vector whose jth element equals one and whose additional elements are zero.
$E_{m\times n}^{(i,j)}$	the $m\times n$ matrix whose (i,j) element equals one and whose additional elements are zero.
$\mathrm{col}_j(S)$	the jth column of the matrix S.
$\mathrm{vec}(\cdot)$	the invertible linear operator defined such that $\mathrm{vec}(S)\triangleq[s_1^Ts_2^T\cdots s_q^T]^T$, $S\in\mathbb{R}^{p\times q}$, where $s_j\triangleq\mathrm{col}_j(S)$.
$\overline{\mathrm{vec}}(\cdot)$	the invertible linear operator (called the "reduced vec operator" [14]) defined such that $\overline{\mathrm{vec}}(S)\triangleq[s_{11};\ s_{12}, s_{22};\ \ldots;\ s_{1p}, s_{2p}, \ldots, s_{pp}]^T$, $S\in\mathbb{S}^{p\times p}$.
$\otimes$	the Kronecker product [3], [9].
$A\leftarrow B$	"A is replaced by B."
$\lambda_i(A)$	the ith eigenvalue of the square matrix A (the eigenvalue ordering can be arbitrary).
$\rho(A)$	the spectral radius of the square matrix A.

1. Introduction.

Lyapunov equations play a fundamental role in linear systems theory. As a result, extensive research has been devoted to developing solution techniques for this class of equations (e.g., [1], [8], [10], [15]). In this paper we consider the problem of finding solutions $Q\in\mathbb{S}^{n\times n}$ to the modified Lyapunov equation

$$0 = AQ + QA^T + \mathscr{F}(Q) + V,\quad A\in\mathbb{R}^{n\times n},\quad V\in\mathbb{S}^{n\times n}, \tag{1.1}$$

* Received by the editors February 25, 1991; accepted for publication (in revised form) September 4, 1991.

† Harris Corporation, Government Aerospace Systems Division, MS 22/4849, Melbourne, Florida 32902 (ecollins@x102a.ess.harris.com, srichter@x102a.ess.harris.com, and ldavis@x102a.ess.harris.com).

where $\lambda_i(A) + \lambda_j(A) \neq 0$ for $i, j \in \{1, \ldots, n\}$ and $\mathscr{F} : \mathbb{S}^{n\times n} \to \mathbb{S}^{n\times n}$ is a linear operator. It will be subsequently shown that although it may not be convenient to do so, for some positive integer p it is always possible to express $\mathscr{F}(Q)$

$$\mathscr{F}(Q) \triangleq \sum_{i=1}^{p} (B_i Q C_i^T + C_i Q B_i^T). \tag{1.2}$$

Linear equations of the form (1.1) have appeared in robustness analysis theory [2] and have also appeared in key substeps of algorithms developed to enable the design of robust, fixed-order control laws [4], [5], [11]–[13].

One technique for solving (1.1) when $\mathscr{F}(Q)$ is given by (1.2) is to express it as the linear vector equation [3], [9]

$$S \operatorname{vec}(Q) = -\operatorname{vec}(V), \tag{1.3}$$

where

$$S \triangleq A \otimes I_n + I_n \otimes A + \sum_{i=1}^{p} (B_i \otimes C_i + C_i \otimes B_i). \tag{1.4}$$

A solution Q is then found by solving the linear vector equation (1.3) in n^2 unknowns. The solution is unique if and only if $S \in \mathbb{R}^{n^2 \times n^2}$ is nonsingular, in which case Q is found by inverting S or by using an alternative solution methodology, such as Gaussian elimination. Obviously, for large n this task can be computationally prohibitive.

Some relief can be found by exploiting the fact that it is known a priori that a solution Q that solves (1.1) is symmetric. In this case, if $\mathscr{F}(Q)$ is defined by (1.2), it is possible to reduce (1.1) to a linear vector equation in $n(n+1)/2$ unknowns. Unfortunately, the size of the matrix to be inverted is still very large for large n.

However, suppose that $\mathscr{F}(Q)$ is in some sense of low rank (to be made precise later). For example, consider the extreme case in which

$$\mathscr{F}(Q) \triangleq EQE, \tag{1.5}$$

where $E \in \mathbb{R}^{n\times n}$ is given by

$$E \triangleq \begin{bmatrix} 1 & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & 0_{(n-1)\times(n-1)} \end{bmatrix}. \tag{1.6}$$

Then, since (1.1) is just a "slight" perturbation from a standard Lyapunov equation, it seems plausible that we can find a computational scheme that requires only slightly more computations than those required for a standard Lyapunov equation.

In what follows, we develop a solution technique for (1.1) that relies on decomposing $\mathscr{F}(Q)$ as $\mathscr{F}(Q) = \mathscr{G}(\phi(Q))$, where $\phi(Q)$ is an m-dimensional vector. This solution methodology does not require directly converting the matrix equation to a linear vector equation, although it does require the solution of a certain linear vector equation of dimension m. Once the solution x of this linear equation is computed, the desired solution Q is found by solving a standard Lyapunov equation. Hence if m is small, the new method for finding Q will have very significant computational advantages over methods based on a straightforward transformation of the modified Lyapunov equation to a linear vector equation.

The paper is organized as follows. In § 2 we lay the theoretical foundation for the solution technique. Section 3 builds on this theoretical foundation and develops a computationally efficient solution procedure. Section 4 illustrates the results by considering two modified Lyapunov equations that occur in engineering practice. The results are extended in § 5 to develop a solution procedure for an arbitrary number of coupled modified Lyapunov equations in which the coupling terms are linear operators. Final remarks are presented in § 6.

2. On the solution of the modified Lyapunov equation. Define the Lyapunov operator $\mathscr{L} : \mathbb{R}^{n\times n} \times \mathbb{S}^{n\times n} \rightarrow \mathbb{S}^{n\times n}$ such that $\mathscr{L}(A, V)$ is the solution of the Lyapunov equation

$$0 = A\mathscr{L}(A, V) + \mathscr{L}(A, V)A^T + V. \tag{2.1}$$

Since it is assumed that $\lambda_i(A) + \lambda_j(A) \neq 0$ for $i, j \in \{1, \ldots, n\}$, it is well known (see, e.g., [1]) that $\mathscr{L}(A, \cdot)$ is invertible. Note that $\mathscr{L}(A, \cdot)$ is also a linear operator.

Lemma 2.1 reveals that the modification term $\mathscr{F}(Q)$ can always be written in a "nice" form.

LEMMA 2.1. *Let $\mathscr{F} : \mathbb{S}^{n\times n} \rightarrow \mathbb{S}^{n\times n}$ be a linear operator. Then, for some integer p, $\mathscr{F}(Q)$ can be expressed as*

$$\mathscr{F}(Q) = \sum_{i=1}^{p} (B_iQC_i^T + C_iQB_i^T), \tag{2.2}$$

where each B_i and C_i is a member of $\mathbb{R}^{n\times n}$.

Proof. The proof is by construction. We begin by decomposing $\mathscr{F}(Q)$ as follows:

$$\begin{aligned}\mathscr{F}(Q) = &\sum_{i=1}^{n} \left[\frac{1}{2}(q_{ii}I_n)\mathscr{F}(E_{n\times n}^{(i,i)}) + \frac{1}{2}\mathscr{F}(E_{n\times n}^{(i,i)})(q_{ii}I_n)\right] \\ &+ \sum_{i=1}^{n}\sum_{j=i+1}^{n} [(q_{ij}I_n)\mathscr{F}(E_{n\times n}^{(i,j)}) + \mathscr{F}(E_{n\times n}^{(j,i)})(q_{ji}I_n)].\end{aligned} \tag{2.3}$$

Now

$$q_{ij}I_n = \sum_{k=1}^{n} E_{n\times n}^{(k,i)}QE_{n\times n}^{(j,k)}. \tag{2.4}$$

By substituting (2.4) into (2.3) and recognizing that $q_{ji} = q_{ij}$, we obtain

$$\begin{aligned}\mathscr{F}(Q) = &\sum_{i=1}^{n}\sum_{k=1}^{n} \left[\frac{1}{2}E_{n\times n}^{(k,i)}QE_{n\times n}^{(i,k)}\mathscr{F}(E_{n\times n}^{(i,i)}) + \frac{1}{2}\mathscr{F}(E_{n\times n}^{(i,i)})E_{n\times n}^{(k,i)}QE_{n\times n}^{(i,k)}\right] \\ &+ \sum_{i=1}^{n}\sum_{j=i+1}^{n}\sum_{k=1}^{n} [E_{n\times n}^{(k,i)}QE_{n\times n}^{(j,k)}\mathscr{F}(E_{n\times n}^{(i,j)}) + \mathscr{F}(E_{n\times n}^{(j,i)})E_{n\times n}^{(k,j)}QE_{n\times n}^{(i,k)}].\end{aligned} \tag{2.5}$$

Since $E_{n\times n}^{(j,i)} = E_{n\times n}^{(i,j)^T}$, the proof is immediate from (2.5). □

Although Lemma 2.1 guarantees that it is always possible to express $\mathscr{F}(Q)$ by a summation of the form (2.2), it may not be computationally efficient to compute the B_i and C_i in (2.2).

For the solution methodology of this paper it is more important to decompose $\mathscr{F}(Q)$ as

$$\mathscr{F}(Q) = \mathscr{G}(\phi(Q)), \tag{2.6}$$

where $\phi : \mathbb{S}^{n\times n} \rightarrow \mathbb{R}^m$ and $\mathscr{G} : \mathbb{R}^m \rightarrow \mathbb{S}^{n\times n}$. As an illustration, if $\mathscr{F}(Q)$ is given by (1.5) and (1.6), then $\phi(\cdot)$ and $\mathscr{G}(\cdot)$ can be chosen such that $\phi(Q) \triangleq q_{11}$ and

$$\mathscr{G}(q_{11}) = \begin{bmatrix} q_{11} & 0_{1\times(n-1)} \\ 0_{(n-1)\times 1} & 0_{(n-1)\times(n-1)} \end{bmatrix}.$$

From (2.6) it follows that any solution Q of (1.1) can be expressed as

$$Q = \mathscr{L}(A, \mathscr{G}(x) + V), \tag{2.7}$$

where

$$x = \phi(Q). \tag{2.8}$$

Thus if x is somehow known, we can find the corresponding Q by simply solving a Lyapunov equation with forcing function $\mathscr{G}(x) + V$. With this in mind we now state a key theorem.

THEOREM 2.2. *Let $M \in \mathbb{R}^{m\times m}$ be defined such that*

$$My \triangleq \phi(\mathscr{L}(A, \mathscr{G}(y))), \qquad y \in \mathbb{R}^m, \tag{2.9}$$

and let

$$d \triangleq \phi(\mathscr{L}(A, V)). \tag{2.10}$$

Then there exists a solution Q to (1.1) *if and only if there exists a solution x to*

$$x = Mx + d. \tag{2.11}$$

In addition, the set of solutions $\mathbf{Q}$ *to* (1.1) *is given by*

$$\mathbf{Q} \triangleq \{Q : Q = \mathscr{L}(A, \mathscr{G}(x) + V),\ x \text{ satisfies } (2.11)\}. \tag{2.12}$$

Proof. Q is a solution of (1.1) if and only if

$$Q = \mathscr{L}(A, \mathscr{G}(\phi(Q)) + V), \tag{2.13}$$

which implies

$$\begin{aligned}\phi(Q) &= \phi(\mathscr{L}(A, \mathscr{G}(\phi(Q)) + V)\\ &= \phi(\mathscr{L}(A, \mathscr{G}(\phi(Q)))) + \phi(\mathscr{L}(A, V)).\end{aligned} \tag{2.14}$$

It follows from (2.9) and (2.10) that (2.14) is equivalent to

$$\phi(Q) = M\phi(Q) + d. \tag{2.15}$$

Thus if (1.1) has a solution Q, there exists a solution x of (2.11).

Next, assume that x is a solution of (2.11) or, equivalently,

$$x = \phi(\mathscr{L}(A, \mathscr{G}(x) + V)), \tag{2.16}$$

and let

$$Q = \mathscr{L}(A, \mathscr{G}(x) + V). \tag{2.17}$$

It follows from substituting (2.17) into the right-hand side of (2.13) and using (2.16) and (2.17) consecutively that

$$\begin{aligned}\mathscr{L}(A, \mathscr{G}(\phi(Q)) + V) &= \mathscr{L}(A, \mathscr{G}(\phi(\mathscr{L}(A, \mathscr{G}(x) + V))) + V)\\ &= \mathscr{L}(A, \mathscr{G}(x) + V)\\ &= Q.\end{aligned}$$

Hence Q given by (2.17) with x given by (2.16) satisfies (2.13). □

An obvious corollary of Theorem 2.2 is as follows.

COROLLARY 2.3. *The modified Lyapunov equation* (1.1) *has a unique solution if and only if* $(I_m - M)$ *is invertible, in which case the solution Q is given by*

$$Q = \mathscr{L}(A, \mathscr{G}((I_m - M)^{-1}d) + V). \tag{2.18}$$

Theorem 2.2 and Corollary 2.3 show that all solutions to (1.1) can be found by solving the linear vector equation (2.11) in m unknowns and then solving a standard $n \times n$ Lyapunov equation (as described by (2.7)). If m is sufficiently small, it would then appear that this solution technique would be more computationally efficient than would solving a linear vector equation in $n(n+1)/2$ unknowns. However, this is only true if the cost of computing the matrix M and vector d, defined respectively by (2.9) and (2.10), is not too large. Section 3 shows that the primary computational burden of constructing M and d is the solution of $m+1$ Lyapunov equations, each having A as the coefficient matrix. If (1.1) is first transformed to a basis in which A is nearly diagonal, the cost of computing M and d is not large.

3. An efficient solution methodology. For $j \in \{0, 1, \ldots, m\}$ define $Q^{(j)}$ to be the solution of the Lyapunov equation

$$0 = AQ^{(j)} + Q^{(j)}A^T + V^{(j)}, \tag{3.1}$$

where

$$V^{(j)} \triangleq \begin{cases} V, & j = 0, \\ \mathscr{G}(e_m^{(j)}), & j = 1, \ldots, m. \end{cases} \tag{3.2}$$

PROPOSITION 3.1. *Let* $M \in \mathbb{R}^{m \times m}$ *and* $d \in \mathbb{R}^m$ *be defined respectively by* (2.9) *and* (2.10). *Then,*

$$\operatorname{col}_j(M) \triangleq \phi(Q^{(j)}), \qquad j = 1, 2, \ldots, m, \tag{3.3}$$

and

$$d = \phi(Q^{(0)}). \tag{3.4}$$

Proof. First, note that

$$Q^{(0)} = \mathscr{L}(A, V), \tag{3.5}$$

$$Q^{(j)} = \mathscr{L}(A, \mathscr{G}(e_m^{(j)})), \qquad j = 1, 2, \ldots, m. \tag{3.6}$$

It follows from (2.9) that

$$\operatorname{col}_j(M) = \phi(\mathscr{L}(A, \mathscr{G}(e_m^{(j)}))). \tag{3.7}$$

Equation (3.3) is then immediate from (3.7) and (3.6), and (3.4) follows immediately from (2.10) and (3.5). □

Proposition 3.1 reveals that M and d can be constructed by solving $m+1$ Lyapunov equations and operating on each solution with the linear operator $\phi(\cdot)$. The $m+1$ Lyapunov equations can be computed with relatively few computations if A is in a "nice" form, such as nearly diagonal or nearly triangular. Hence transformation into an appropriate basis is the first step in the solution procedure detailed below. The real Schur form [7] (a nearly triangular form) has been shown to be numerically stable and can be chosen as this basis. However, the tridiagonal basis discussed in [6] and [15] is also numerically stable, and when it is used with the alternating-direction implicit algorithm [15] for solving Lyapunov equation it can provide substantial savings in computation time. In particular, it follows from the analysis presented in [15] that if the Schur form were used, the algorithm would require approximately $\frac{41}{3}n^3 + (m+1)n^3$ flops. Here we are ignoring operations involving the transformation of the operator $\mathscr{F}(\cdot)$ that are problem dependent. The first term in the operation count ($\frac{41}{3}n^3$) represents the operations involved with the transformation of the system, whereas the second term ($(m+1)n^3$) represents the operations involved with solving the Lyapunov equations in the Schur basis. The approximate algorithm requirements for the tridiagonal basis are $\frac{19}{3}n^3 + (m+1)12Jn^2$ flops, where J represents the number of iterations of the alternating-direction implicit algorithm. The results of [15] indicate that typical values of J are in the interval [4, 8], so for large n transformation to tridiagonal basis can be much cheaper than transformation to Schur form.

Before summarizing the solution methodology we note that it may be possible to solve for x directly (and perhaps more efficiently) without explicitly calculating M. For example, if $\rho(M) < 1$, then x may be computed [7] by the series

$$x = d + Md + M^2d + M^3d + \cdots$$

or, equivalently,

$$x = d_0 + d_1 + d_2 + d_3 + \cdots,$$

where, as per (2.9), $d_0, d_1, \cdots$ can be found by using

$$d_{i+1} = \phi(\mathscr{L}(A, \mathscr{G}(d_i))), \qquad d_0 = d.$$

This is a subject for future research.

SOLUTION METHODOLOGY

Step 1. Compute a transformation matrix T such that $T^{-1}AT$ is in real Schur form or tridiagonal form.

Step 2. Transform (1.1) to the new basis by making the replacements

$$A \leftarrow T^{-1}AT, \tag{3.8}$$

$$V \leftarrow T^{-1}VT^{-T}, \tag{3.9}$$

$$\mathscr{F}(\cdot) \leftarrow T^{-1}\mathscr{F}(\mathscr{T}(\cdot))T^{-T}, \tag{3.10}$$

where $\mathscr{T} : \mathbb{S}^{n\times n} \rightarrow \mathbb{S}^{n\times n}$ is defined such that

$$\mathscr{T}(Q) \triangleq TQT^T, \qquad Q \in \mathbb{S}^{n\times n}. \tag{3.11}$$

Note that the transformations (3.8)–(3.10) imply

$$Q \leftarrow T^{-1}QT^{-T}. \tag{3.12}$$

Step 3. Define $\phi : \mathbb{S}^{n\times n} \rightarrow \mathbb{R}^m$ and $\mathscr{G} : \mathbb{R}^m \rightarrow \mathbb{S}^{n\times n}$ such that

$$\mathscr{F}(\cdot) = \mathscr{G}(\phi(\cdot)). \tag{3.13}$$

(The objective here is to define $\phi(\cdot)$ and $\mathscr{G}(\cdot)$ such that m is as small as possible.)

Step 4. Construct M and d by using the following loop.

```
FOR $j = 0, m$
    Compute $V^{(j)}$ by using (3.2).
    Compute the solution $Q^{(j)}$ of (3.1).
    If ($j = 0$), compute $d$ by using (3.4).
    If ($j \geq 1$), compute $\text{col}_j(M)$ by using (3.3).
END FOR
```

Step 5. Find a solution x of

$$x = Mx + d. \tag{3.14}$$

If a solution does not exist, then no solution exists to (1.1); so stop.

Step 6. Compute a solution Q to

$$0 = AQ + QA^T + \mathscr{G}(x) + V. \tag{3.15}$$

Step 7. Transform Q back to the original basis, i.e.,

$$Q \leftarrow TQT^T. \tag{3.16}$$

Step 8. Stop.

4. Two illustrations from engineering applications. This section considers two modified Lyapunov equations of the form (1.1) that occur in engineering applications and shows how to construct the $\phi(\cdot)$ and $\mathscr{G}(\cdot)$ in (2.2) to be used in the solution scheme of the previous section.

Example 4.1. This example considers the modified Lyapunov equation

$$0 = AQ + QA^T + H_L S_L Q S_R H_R + H_R^T S_R^T Q S_L^T H_L^T + V, \tag{4.1}$$

where $H_L \in \mathbb{R}^{n \times m_s}$, $S_L \in \mathbb{R}^{m_s \times n}$, $S_R \in \mathbb{R}^{n \times n_s}$, $H_R \in \mathbb{R}^{n_s \times n}$ with $m_s < n$ and $n_s < n$. This equation occurs in an algorithm for reduced-order-controller design [4], [5], [11]. Notice that the structure of (4.1) does not change when it is transformed such that $A \leftarrow T^{-1}AT$ for some nonsingular T. Thus without loss of generality we can assume that A is in real Schur form or tridiagonal form.

Equation (4.1) is equivalent to (1.1) with

$$\mathscr{F}(Q) = H_L S_L Q S_R H_R + H_R^T S_R^T Q S_L^T H_L^T. \tag{4.2}$$

We can express (4.2) as

$$\mathscr{F}(Q) = \mathscr{G}(\phi(Q)), \tag{4.3}$$

where $\phi(\cdot)$ and $\mathscr{G}(\cdot)$ are the linear operators defined such that

$$\phi(Q) \triangleq \operatorname{vec}(S_L Q S_R), \qquad Q \in \mathbb{S}^{n \times n}, \tag{4.4}$$

$$\mathscr{G}(x) \triangleq H_L \operatorname{vec}^{-1}(x) H_R + H_R^T[\operatorname{vec}^{-1}(x)]^T H_L^T, \qquad x \in \mathbb{R}^{m_s n_s}. \tag{4.5}$$

Thus for

$$m = m_s n_s \tag{4.6}$$

$\phi(\cdot)$ maps $\mathbb{S}^{n \times n}$ into $\mathbb{R}^m$ and $\mathscr{G}(\cdot)$ maps $\mathbb{R}^m$ into $\mathbb{S}^{n \times n}$.

Recall that we have assumed that both $m_s < n$ and $n_s < n$. Hence

$$m_s \ll n \quad \text{or} \quad n_s \ll n \Rightarrow m \ll n(n+1)/2, \tag{4.7}$$

and the solution procedure of the previous section will be more efficient than procedures that directly convert (4.1) to a linear vector equation. In the algorithm of [4], [5], and [11] for reduced-order-controller design, $m_s \ll n$ if the order of the desired controller is much less than the order of the design plant and $n_s \ll n$ if the number of sensors (or actuators) is much less than the order of the design plant.

Example 4.2. This example considers the modified Lyapunov equation

$$0 = AQ + QA^T + S\left(\sum_{i=1}^{p} A_i R Q R^T A_i^T\right)S^T + V, \tag{4.8}$$

where $S, T \in \mathbb{R}^{n \times n}$, $p < n$, and for some integer vector ℓ of dimension p the A_i are defined such that

$$A_i = \text{block-diag}\left\{0, \ldots, 0, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, 0, \ldots, 0\right\}, \tag{4.9}$$

where the (ℓ_i, ℓ_i) element of A_i corresponds to the $(1, 1)$ element of

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Equation (4.8) occurs in an algorithm for robust-controller design for flexible structures [4], [5]. As in the previous example, the structure of (4.8) does not change when it is transformed such that $A \leftarrow T^{-1}AT$ for some nonsingular T. Thus without loss of generality we can assume that A is in a modal basis.

Equation (4.7) is equivalent to (1.1) with

$$(4.10) \qquad \mathscr{F}(Q) = S\left(\sum_{i=1}^{p} A_iRQR^TA_i^T\right)S^T.$$

Define $\mathscr{P}_j : \mathbb{S}^{n\times n} \to \mathbb{S}^{2\times 2}$ such that

$$(4.11) \qquad \mathscr{P}_j(W) \triangleq \begin{bmatrix} w_{jj} & w_{j,j+1} \\ w_{j,j+1} & w_{j+1,j+1} \end{bmatrix}, \qquad W \in \mathbb{S}^{n\times n},$$

and define $\mathscr{R} : \mathbb{R}^{3p} \to \mathbb{S}^{n\times n}$ such that

$$(4.12) \qquad \mathscr{R}(x) \triangleq \text{block-diag}\left\{\begin{bmatrix} x_{3(i-1)+1} & x_{3(i-1)+2} \\ x_{3(i-1)+2} & x_{3(i-1)+3} \end{bmatrix}\right\}_{i=1}^{p}, \qquad x \in \mathbb{R}^{3p}.$$

Then (4.10) can be given by

$$(4.13) \qquad \mathscr{F}(Q) = \mathscr{G}(\phi(Q)),$$

where $\phi(\cdot)$ and $\mathscr{G}(\cdot)$ are the linear operators defined such that

(4.14)

$$\phi(Q) \triangleq \text{vec}\,[\overline{\text{vec}}\,(\mathscr{P}_{\ell_1}(A_1TQT^TA_1^T)), \ldots, \overline{\text{vec}}\,(\mathscr{P}_{\ell_p}(A_pTQT^TA_p^T))], \qquad Q \in \mathbb{S}^{n\times n},$$

$$(4.15) \qquad \mathscr{G}(x) \triangleq S\mathscr{R}(x)S^T, \qquad x \in \mathbb{R}^{3p}.$$

Thus for

$$(4.16) \qquad m = 3p,$$

$\phi(\cdot)$ maps $\mathbb{S}^{n\times n}$ into $\mathbb{R}^m$ and $\mathscr{G}(\cdot)$ maps $\mathbb{R}^m$ into $\mathbb{S}^{n\times n}$.

In the algorithm of [4] and [5] for robust-controller design for flexible structures, p represents the number of uncertain modes and hence $p < n$. Thus

$$(4.17) \qquad n \gg 3 \Rightarrow m \ll n(n+1)/2.$$

The condition $n \gg 3$ is equivalent to the order of the design plant being much greater than 3.

5. Extensions to systems of coupled modified Lyapunov equations. In this section we consider the problem of finding a solution $(Q_1, \ldots, Q_r)$ to the coupled modified Lyapunov equations

$$(5.1) \qquad \begin{aligned} 0 &= A_1Q_1 + Q_1A_1^T + \mathscr{F}_1(Q_1, \ldots, Q_r) + V_1, \\ 0 &= A_2Q_2 + Q_2A_2^T + \mathscr{F}_2(Q_1, \ldots, Q_r) + V_2, \\ &\vdots \\ 0 &= A_rQ_r + Q_rA_r^T + \mathscr{F}_r(Q_1, \ldots, Q_r) + V_r, \end{aligned}$$

where, for $k \in \{1, \ldots, r\}$, $A_k \in \mathbb{R}^{n\times n}$ and $V_k \in \mathbb{S}^{n\times n}$. It is also assumed that, for $k \in \{1, \ldots, r\}$, $\mathscr{F}_k : X_1^r \mathbb{S}^{n\times n} \rightarrow \mathbb{S}^{n\times n}$ is a linear operator and $\lambda_i(A_k) + \lambda_j(A_k) \neq 0$ for $i, j \in \{1, \ldots, n\}$. Systems of equations of the form (5.1) must be solved in algorithms developed to enable the design of robust, reduced-order controllers [4], [11], [12].

The results of this section parallel those of §§ 2 and 3. Hence the solution methodology is based on decomposing the $\mathscr{F}_k(\cdot)$ as

$$\mathscr{F}_k(\cdot) = \mathscr{G}_k(\phi(\cdot)), \qquad k = 1, 2, \ldots, m, \tag{5.2}$$

where $\phi : \mathbb{S}^{n\times n} \rightarrow \mathbb{R}^m$ and each $\mathscr{G}_k : \mathbb{R}^m \rightarrow \mathbb{S}^{n\times n}$ is a linear operator.

Lemma 5.1 below generalizes Lemma 2.1 and reveals that each of the modification terms $\mathscr{F}_k(\cdot)$ can be written in a "nice" form.

LEMMA 5.1. *Let $\mathscr{F}_k : X_1^r \mathbb{S}^{n\times n} \rightarrow \mathbb{S}^{n\times n}$ be a linear operator. Then, for some set of integers $\{p_i\}_{i=1}^r$, $\mathscr{F}_k(Q_1, \ldots, Q_r)$ can be expressed as*

$$\begin{aligned}\mathscr{F}_k(Q_1, \ldots, Q_r) = &\sum_{i=1}^{p_1} (B_{1,i}Q_1C_{1,i}^T + C_{1,i}QB_{1,i}^T) \\ &+ \cdots + \sum_{i=1}^{p_r} (B_{p,i}Q_rC_{p,i}^T + C_{p,i}QB_{p,i}^T),\end{aligned} \tag{5.3}$$

where each $B_{k,i}$ and $C_{k,i}$ is a member of $\mathbb{R}^{n\times n}$.

Proof. The proof is essentially identical to the proof of Lemma 2.1 and is thus omitted. □

THEOREM 5.2. *Let $M \in \mathbb{R}^{m\times m}$ be defined such that*

$$My \triangleq \phi(\mathscr{L}(A_1, \mathscr{G}_1(y)), \ldots, \mathscr{L}(A_r, \mathscr{G}_r(y))), \qquad y \in \mathbb{R}^m, \tag{5.4}$$

and let $d \in \mathbb{R}^m$ be defined by

$$d \triangleq \phi(\mathscr{L}(A_1, V_1), \ldots, \mathscr{L}(A_r, V_r)). \tag{5.5}$$

Then (5.1) has a solution $(Q_1, \ldots, Q_r)$ if and only if there exists a solution x to

$$x = Mx + d. \tag{5.6}$$

In addition, the set of solutions $\mathbf{Q}^r$ to (5.1) is given by

$$\begin{aligned}\mathbf{Q}^r = \{(Q_1, \ldots, Q_r) : Q_k = \mathscr{L}(A_k, \mathscr{G}_k(x) + V_k)& \\ \textit{for } k = 1, \ldots, r; x \textit{ satisfies } (5.6)\}&.\end{aligned} \tag{5.7}$$

Proof. $(Q_1^r, \ldots, Q_r)$ is a solution of (5.1) if and only if

$$Q_k = \mathscr{L}(A_k, \mathscr{G}_k(\phi(Q_1, \ldots, Q_r)) + V_k), \qquad k = 1, \ldots, r, \tag{5.8}$$

which implies that

$$\begin{aligned}\phi(Q_1, \ldots, Q_r) = \phi(&\mathscr{L}(A_1, \mathscr{G}_1(\phi(Q_1, \ldots, Q_r)) + V_1), \ldots, \\ &\mathscr{L}(A_r, \mathscr{G}_r(\phi(Q_1, \ldots, Q_r)) + V_r))\end{aligned} \tag{5.9}$$

or, equivalently,

(5.10)

$$\begin{aligned}\phi(Q_1, \ldots, Q_r) = \phi(\mathscr{L}(A_1, \mathscr{G}_1(\phi(Q_1, \ldots, Q_r))), \ldots, \mathscr{L}(A_r, \mathscr{G}_r(\phi(Q_1, \ldots, Q_r)))) \\ + \phi_k(\mathscr{L}(A_1, V_1), \ldots, \mathscr{L}(A_r, V_r)).\end{aligned}$$

It follows from (5.4) and (5.5) that (5.10) is equivalent to

$$\phi(Q_1, \ldots, Q_r) = M\phi(Q_1, \ldots, Q_r) + d. \tag{5.11}$$

Thus if (5.1) has a solution, there exists a solution x of (5.6).

Next, assume that x is a solution of (5.6) or, equivalently,

$$x = \phi(\mathscr{L}(A_1, \mathscr{G}_1(x_1) + V_1), \ldots, \mathscr{L}(A_r, \mathscr{G}_r(x_r) + V_r)), \qquad k = 1, \ldots, r, \tag{5.12}$$

and that $(Q_1, \ldots, Q_r)$ is given by

$$Q_k = \mathscr{L}(A_k, \mathscr{G}_k(x) + V_k), \qquad k = 1, \ldots, r. \tag{5.13}$$

It follows by substituting (5.13) into the right-hand side of (5.8) and using (5.12) and (5.13) consecutively that

$$\begin{aligned}
&\mathscr{L}_k(A_k, \mathscr{G}_k(\phi(Q_1, \ldots, Q_r)) + V_k)\\
&\quad = \mathscr{L}_k(A_k, \mathscr{G}_k(\phi(\mathscr{L}(A_1, \mathscr{G}_1(x) + V_1), \ldots, \mathscr{L}(A_r, \mathscr{G}_r(x) + V_r)) + V_k))\\
&\quad = \mathscr{L}_k(A_k, \mathscr{G}_k(x) + V_k)\\
&\quad = Q_k.
\end{aligned}$$

□

COROLLARY 5.3. *The system of coupled Lyapunov equations* (5.1) *has a unique solution if and only if* $(I_m - M)$ *is invertible, in which case the solution* $(Q_1, \ldots, Q_r)$ *is given by*

$$Q_k = \mathscr{L}_k(A_k, \mathscr{G}((I - M)^{-1}d) + V), \qquad k = 1, \ldots, r. \tag{5.14}$$

Now, for $j = \{0, 1, \ldots, m\}$ define the matrices $Q_i^{(j)}$ to be the solutions of the Lyapunov equations

$$0 = A_i Q_i^{(j)} + Q_i^{(j)} A_i^T + V_i^{(j)}, \qquad i = 1, \ldots, r, \tag{5.15}$$

where

$$V_i^{(j)} = \begin{cases} V_i, & j = 0\\ \mathscr{G}_i(e_m^{(j)}), & j = 1, \ldots, m_i. \end{cases} \tag{5.16}$$

PROPOSITION 5.4. *Let* $M \in \mathbb{R}^{m \times m}$ *and* $d \in \mathbb{R}^m$ *be defined respectively by* (5.4) *and* (5.5). *Then*

$$\operatorname{col}_j(M) \triangleq \sum_{i=1}^{r} \phi(0, \ldots, 0, Q_i^{(j)}, 0, \ldots, 0) \tag{5.17}$$

and

$$d \triangleq \sum_{i=1}^{r} \phi(0, \ldots, 0, Q_i^{(0)}, 0, \ldots, 0), \tag{5.18}$$

where $Q_i^{(j)}$ and $Q_i^{(0)}$ in, respectively, (5.17) and (5.18) correspond to the ith argument of the function $\phi(\cdot)$.

Proof. First, note that for $i \in \{1, \ldots, r\}$

$$Q_i^{(0)} = \mathscr{L}(A_i, V_i), \tag{5.19}$$

$$Q_i^{(j)} = \mathscr{L}(A_i, \mathscr{G}_i(e_m^{(j)})), \qquad j = 1, \ldots, m. \tag{5.20}$$

It follows from (5.4) that

$$\text{col}_j\,(M) = \phi(\mathscr{L}(A_1, \mathscr{G}_1(e_m^{(1)})), \ldots, \mathscr{L}(A_r, \mathscr{G}_r(e_m^{(r)}))). \tag{5.21}$$

Equation (5.17) is then immediate from (5.21) and (5.20), and (5.18) follows immediately from (5.5) and (5.19). □

Proposition 5.4 reveals that M and d, defined respectively by (5.4) and (5.5), can be constructed by solving $(m+1)r$ Lyapunov equations or, specifically, $m+1$ Lyapunov equations with coefficient matrix A_i for $i = 1, \ldots, r$. These Lyapunov equations can be computed with relatively few computations if the A_i are in real Schur form or tridiagonal form. Hence transformation into an appropriate basis is the first step in the solution procedure detailed below.

Step 1. For $i = 1, \ldots, r$ compute a transformation matrix T_i such that $T_i^{-1}A_iT_i$ is in real Schur form or tridiagonal form.

Step 2. Transform the equation (5.1) by making the following replacements.

For $i = 1, \ldots, r$ let

$$A_i \leftarrow T_i^{-1}A_iT_i, \tag{5.22}$$

$$V_i \leftarrow T_i^{-1}V_iT_i^{-T}, \tag{5.23}$$

$$\mathscr{F}_i(\cdot, \ldots, \cdot) \leftarrow T_i^{-1}\mathscr{F}(\mathscr{T}_1(\cdot), \ldots, \mathscr{T}_r(\cdot))T_i^{-T}, \tag{5.24}$$

where $\mathscr{T}_i : \mathbb{S}^{n\times n} \to \mathbb{S}^{n\times n}$ is defined such that

$$\mathscr{T}_i(Q) \triangleq T_iQT_i^T, \qquad Q \in \mathbb{S}^{n\times n}. \tag{5.25}$$

Note that the transformations (5.22)–(5.25) imply

$$Q_i \leftarrow T_i^{-1}Q_iT_i^{-T}. \tag{5.26}$$

Step 3. For $i = 1, \ldots, r$ define $\phi_i : \mathbb{S}^{n\times n} \to \mathbb{R}^{m_i}$ and $\mathscr{G}_i : \mathbb{R}^{m_i} \to \mathbb{S}^{n\times n}$ such that

$$\mathscr{F}_i(\cdot) = \mathscr{G}_i(\phi_i(\cdot)). \tag{5.27}$$

(The objective here is to define $\phi_i(\cdot)$ and $\mathscr{G}_i(\cdot)$ such that m_i is as small as possible.)

Step 4. Construct M and d by using the following loops.

Initialize $d = 0$.

Initialize $\mathscr{M}_j = 0$ for $j = 1, \ldots, m$.

```
FOR $i = 1, r$,
    FOR $j = 0, m$
        Compute $V_i^{(j)}$ by using (5.16).
        Compute the solution $Q_i^{(j)}$ of (5.15).
        If ($j = 0$), compute $d$ by using (5.18).
        If ($j \geqq 1$), compute $d$ by using (5.17).
    END FOR
END FOR
```

Step 5. Find a solution x of

$$x = Mx + d. \tag{5.28}$$

If a solution does not exist, then no solution exists to (5.1); so stop.

Step 6. For $k = 1, \ldots, r$ compute the solution Q_k of

$$A_kQ_k + Q_kA_k^T + \mathscr{G}_k(x) + V_k. \tag{5.29}$$

Step 7. For $k = 1, \dots, r$ transform the Q_k back to the original basis, i.e.,

$$Q_k \leftarrow T_k Q_k T_k^T. \tag{5.30}$$

Step 8. Stop.

6. Final remarks. This paper has presented a procedure for solving a class of modified Lyapunov equations that relies on decomposing the additional term $\mathscr{F}(Q)$ in the Lyapunov equation as $\mathscr{F}(Q) = \mathscr{G}(\phi(Q))$, where $\phi(Q) \in \mathbb{R}^m$. If m is small, the solution procedure will not require much more computation than that required for a standard Lyapunov equation and can thus be much more efficient than a solution procedure based on converting the modified Lyapunov equation to a linear vector equation. The results for the solutions of a single modified Lyapunov equation have also been extended to develop a solution methodology for an arbitrary number of coupled Lyapunov equations in which the coupling terms are linear operators.

Acknowledgments. The authors thank Dr. Dennis Bernstein for several helpful suggestions and comments. They also thank an anonymous reviewer for suggesting the direct solution for x described in § 3.

REFERENCES

[1] R. H. BARTELS AND G. W. STEWART, *Solution of the matrix equation $AX + XB = C$*, Comm. ACM, 15 (1972), pp. 820–826.

[2] D. S. BERNSTEIN AND W. M. HADDAD, *Robust stability and performance via fixed-order dynamic compensation with guaranteed cost bounds*, Math. Control Signals Systems, 3 (1990), pp. 139–163.

[3] J. W. BREWER, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 772–781.

[4] E. G. COLLINS, JR. AND S. RICHTER, *A homotopy algorithm for synthesizing fixed-order robust controllers for flexible structures via the maximum entropy design equations*, in Third Air Force/NASA Symposium on Recent Advances in Multidisciplinary Analysis and Optimization, San Francisco, CA, September 1990.

[5] E. G. COLLINS, JR., S. RICHTER, AND L. D. DAVIS, *A homotopy algorithm for robust, reduced order compensator design via the maximum entropy-optimal projection equations*, Report, Sandia National Laboratories, Albuquerque, NM, 1990.

[6] G. A. GEIST, A. LU, AND E. L. WACHSPRESS, *Stabilized Gaussian reduction of an arbitrary matrix to tridiagonal form*, Tech. Rep. ORNL/TM-11089, Oak Ridge National Laboratory, Oak Ridge, TN, 1989.

[7] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, John Hopkins University Press, Baltimore, MD, 1989.

[8] A. S. HODEL AND K. POOLLA, *Parallel algorithms for the solution of large Lyapunov equations,* SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1189–1203.

[9] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Academic Press, New York, 1985.

[10] B. P. MOLINARI, *Algebraic solution of matrix linear equations in control theory*, Proc. IEE, 116 (1969), pp. 1748–1754.

[11] S. RICHTER, *A homotopy algorithm for solving the optimal projection equations for fixed-order dynamic compensation: Existence, convergence and global optimality*, in Proc. American Controls Conference, Minneapolis, MN, June 1987, pp. 1527–1531.

[12] S. RICHTER AND E. G. COLLINS, JR., *A homotopy algorithm for reduced-order controller design using the optimal projection equations*, in Proc. IEEE Conference on Decision and Control, Tampa, FL, December 1989, Institute of Electrical and Electronics Engineers, New York, 1989, pp. 506–511.

[13] S. RICHTER AND A. S. HODEL, *Homotopy methods for the solution of general modified algebraic Riccati equations*, in Proc. IEEE Conference on Decision and Control, Honolulu, HI, December 1990, Institute of Electrical and Electronics Engineers, New York, 1990, pp. 971–976.

[14] W. J. Vetter, *Vector structures and solutions of linear matrix equations*, Linear Algebra Appl., 10 (1975), pp. 181–188.

[15] A. LU AND E. L. WACHSPRESS, *Solution of Lyapunov equations by alternating direction implicit iteration*, Comput. Math. Appl., 21 (1991), pp. 43–58.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 432–455, April 1993

009

INCREMENTAL UNKNOWNS IN FINITE DIFFERENCES: CONDITION NUMBER OF THE MATRIX*

MIN CHEN† AND ROGER TEMAM‡

Abstract. The utilization of incremental unknowns (IU) with multilevel finite differences was proposed in [R. Temam, *SIAM J. Math. Anal.*, 21 (1991), pp. 154–178] for the integration of elliptic partial differential equations, instead of the usual nodal unknowns. Although turbulence and nonlinear problems were the primary motivations, it appears that the IU method is also interesting for linear problems. For such problems it was shown in [M. Chen and R. Temam, *Numer. Math.*, 59 (1991), pp. 255–271] that the incremental unknown method which is very easy to program is also very efficient, in fact, it is comparable to the classical V-cycle multigrid method.

In this article the condition number of the five-points discretization matrix in space dimension two for the Dirichlet problem is analyzed; more general second-order elliptic boundary value problems are also considered. It is shown that the condition number is $O((\log h)^2)$ where h is the mesh size instead of $O(1/h^2)$ with the usual nodal unknowns. This gives a theoretical justification of the efficiency of the method since the number of operations needed to solve the linear system by the conjugate gradient methods is $O(\sqrt{\kappa})$, where κ is the condition number of the matrix.

Key words. finite differences, incremental unknowns, multigrid methods, linear algebra

AMS(MOS) subject classifications. 65F10, 65F35, 65N20, 35A40

1. Introduction.

After spatial discretization by finite differences, the solution of an elliptic boundary value problem reduces to the solution of a linear system

$$AU = b, \tag{1.1}$$

where $U, b \in \mathbb{R}^N$ and A is a regular matrix of order N. The dimension N is the number of nodal points and U is the vector corresponding to the nodal values of the unknown function. When N is very large, e.g., when nonlinear phenomena and turbulence are involved, we have advocated in [T2] the utilization of the incremental unknowns (IUs) whose definition is recalled below. Let $\bar{U} \in \mathbb{R}^N$ be the vector corresponding to the IUs and let S be the transfer matrix

$$U = S\bar{U}. \tag{1.2}$$

Then (1.1) becomes

$$AS\bar{U} = b$$

or

$$\bar{A}\bar{U} = \bar{b} \tag{1.3}$$

with $\bar{A} = {}^tSAS$, $\bar{b} = {}^tSb$.

IUs can be defined when multilevel discretizations are used. For instance, if two levels of discretizations are used, we have a coarse grid of mesh $2h$ and a fine grid of mesh h; then the incremental unknowns are:

* Received by the editors January 7, 1991; accepted for publication (in revised form) October 7, 1991. This work was partly supported by National Science Foundation grant NSF-DMS8802596, by United States Navy grant N00014-91-5-1140, and by the Research Fund of Indiana University.

† Institute for Scientific Computing and Applied Mathematics, Indiana University, Bloomington, Indiana 47405 (chenm@math.psu.edu).

‡ Laboratoire d'Analyse Numérique, Université Paris-Sud, Bâtiment 425, 91405 Orsay, France and Institute for Applied Mathematics and Scientific Computing, Indiana University, Bloomington, Indiana 47405.

- the nodal values of the unknown function at the coarse grid points, and
- the increment to the averaged value of the function at the closest coarse grid points, for the fine grid points (see Fig. 1.1 and the more precise definition below).

In that case, the matrix S is very simple but $\bar{A} = {}^tSAS$ is complicated, essentially a full matrix [CT1]. However for the solution of (1.3) we can use a conjugate gradient method, in which case we never need the explicit form of $\bar{A}$ since we must only perform the product of $\bar{A}$ with a vector of $\mathbb{R}^N$; this can be done easily and necessitates a small number of algebraic operations.

For example, in space dimension one, let Ω be the interval $(0, 1)$, and let $h = 1/2N$, $N \in \mathbb{N}$ be the fine mesh, and $2h$ be the coarse mesh. If $u_j \simeq u(jh)$ is the approximate value of the unknown function u at the mesh point jh, then the incremental unknowns are:

- $y_{2i} = u_{2i}$ at the coarse grid points $2ih$, $i = 0, \dots, N$;
- $z_{2i+1} = u_{2i+1} - \frac{1}{2}(u_{2i} + u_{2i+2})$ at the fine grid points $(2i + 1)h$, $i = 0, \dots, N - 1$.

Similarly, in space dimension two, let Ω be the square $(0, 1)^2$, and let $h = 1/2N$, $N \in \mathbb{N}$, and $2h$ be the fine mesh and the coarse mesh. If $u_{ij} \simeq u(ih, jh)$ is the approximate value of the unknown function u at the mesh point (ih, jh), we then define the incremental unknowns as follows:

- $y_{2i,2j} = u_{2i,2j}$ at the coarse grid points $(2ih, 2jh)$, $i, j = 0, \dots, N$;
- $z_{2i,2j+1} = u_{2i,2j+1} - \frac{1}{2}(u_{2i,2j} + u_{2i,2j+2})$ at the fine grid point $(2ih, (2j + 1)h)$ between two vertical coarse grid points, $i = 0, \dots, N, j = 0, \dots, N - 1$;
- $z_{2i+1,2j} = u_{2i+1,2j} - \frac{1}{2}(u_{2i,2j} + u_{2i+2,2j})$ at a fine grid point $((2i + 1)h, 2jh)$ between horizontal grids points, $i = 0, \dots, N - 1, j = 0, \dots, N$;
- $z_{2i+1,2j+1} = u_{2i+1,2j+1} - \frac{1}{4}(u_{2i,2j} + u_{2i+2,2j} + u_{2i,2j+2} + u_{2i+2,2j+2})$ at a fine grid point $((2i + 1)h, (2j + 1)h)$ in the center of a coarse grid square, $i, j = 0, \dots, N - 1$.

We can also use several levels of discretizations with meshes $h = 1/(N \cdot 2^j)$, $j = 0, \dots, d$, $N \in \mathbb{N}$. In this case the definition is the same as above at each level of mesh refinement, starting from the finest grid. The reader is referred to [CT1] for the notations and the matrix structure of the problem in this case, while the functional aspect of the problem is studied below.

Although it has its own originality, the IU method can be compared to other existing methods. First, in the language of linear algebra, it amounts to a linear change of variables and therefore to a preconditioner (see, e.g., [AB] or [S]). In the very simple case of space dimension one, it is equivalent to the cyclic reduction method of Golub and Van Loan [GV], Golub and Meurant [GM], and Birkhoff, Varga, and Young [BVY]. Like the multigrid method, the IU method is based on the utilization of several nested grids in finite differences. However, the resolution of the linear system seems different: the unknowns remain the nodal values when the multigrid method is used, and we make use here of the conjugate gradient method, which is not usually the case with multigrid methods.

The IU method presented here can also be compared to the multilevel methods associated with the utilization of hierarchical basis multigrid preconditioners in finite

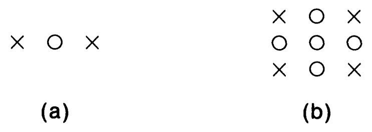


FIG. 1.1. *Coarse grid points* (×) *and fine grid points* (○) *in space dimensions* 1 (a) *and* 2 (b).

elements (see, e.g., Yserentant [Y1], [Y2]; Dryja and Widlund [DW]; Xu [X1], [X2], [X3]; and Atanga and Silvester [AS]). In fact, the incremental unknowns are the same as the components of the function in the Q_1-hierarchical basis and subsequently we use results from [Y1]. However the IU method is developed strictly in the context of finite differences. It is simple to implement and it does not carry the drawbacks (or advantages) of finite element methods. Also, we will present in a subsequent paper several different definitions of incremental unknowns which are not associated with finite elements (see [CT2]). Finally let us recall that the primary motivation and originality of the IU method lies in its utilization for nonlinear problems (see [T2] and articles to appear), but we believe it is also worth exploring its validity in the context of linear problems.

Our aim in this article is to study the condition number of the transformed matrix $\bar{A}$ for second-order elliptic boundary value problems when incremental unknowns are used. In particular, we want to show that the condition number is of order d^{-2} when d levels of discretization are used and the finest mesh is of order $h(=h_d) = h_0/2^d$. Hence the condition number is $O((\log h)^2)$, contrarily to the condition number of A, which is $O(h^{-2})$. This result is obtained by deriving appropriate bounds on the smallest and largest eigenvalues of $\bar{A}$ through the associated bilinear form. The considerable decrease in the condition number occurring in finite differences with the utilization of the IUs can be compared to that occurring in finite element methods when multilevel discretizations and hierarchical bases are used (see [Y1]).

This article is organized as follows. In § 2 we describe the mathematical setting and formulate the main results, namely, the appropriate bounds on the eigenvalues. These results are proved in §§ 3 and 4. Then in § 5 we apply these results to some specific boundary value problems.

2. Mathematical setting. The main results. We consider in $\mathbb{R}^2$ two discretization meshes

$$h_0 = (h_{1,0}, h_{2,0}), \quad h_d = (h_{1,d}, h_{2,d}), \quad h_{i,d} = \frac{h_{i,0}}{2^d}, \quad d > 0;$$

h_0 is the coarse mesh and h_d the fine mesh and different meshes are allowed in both directions x_1, x_2. To these meshes we associate the grids

$$\mathcal{R}_0 = \mathcal{R}_{h_0} \quad \text{made of points } (j_1 h_{1,0}, j_2 h_{2,0}),$$

$$\mathcal{R}_d = \mathcal{R}_{h_d} \quad \text{made of points } (j_1 h_{1,d}, j_2 h_{2,d}),$$

where $j_1, j_2 \in \mathbb{Z}$. For $j = (j_1, j_2) \in \mathbb{Z}^2$ we denote also by K_{j,h_d}, or more simply $K_{j,d}$, the rectangle

$$K_{j,d} = (j_1 h_{1,d}, (j_1 + 1) h_{1,d}) \times (j_2 h_{2,d}, (j_2 + 1) h_{2,d}).$$

We define in a similar manner a rectangle $K_{j,0}$ or K_{j,h_0} by replacing h_d by h_0.

For simplicity we shall emphasize the case where Ω is a rectangle $(0, a_1) \times (0, a_2)$ and $h_{i,0} = a_i/N$, $i = 1, 2$. More general open sets Ω and more general meshes will be considered below. Although we are interested in *finite differences*, a space of finite element functions on Ω will be used. More precisely, we denote by V_{h_d}, or more simply by V_d, the space of continuous real functions on $\bar{\Omega}$ that are Q_1§ on each rectangle $K_{j,d} \subset \Omega$, and by $\mathcal{U}_d$ the set of nodal points

$$\mathcal{U}_d = \mathcal{R}_d \cap \bar{\Omega},$$

§ That is, the function is affine with respect to x_1 and x_2 separately.

and we recall that the functions of V_d are uniquely defined by their nodal values, i.e., their values on $\mathscr{U}_d$. We define in a similar manner the spaces $V_l\,(=V_{hl})$, $0 \leqq l \leqq d$ and observe that

$$V_0 \subset V_1 \subset \cdots \subset V_d. \tag{2.1}$$

In addition, V_d and the subspaces V_l are included in the Sobolev space $H^1(\Omega)$, and we can endow V_d with the scalar products and norms induced by $L^2(\Omega)$ and $H^1(\Omega)$:

$$(u, v) = \int_\Omega u(x)v(x)\,dx, \quad |u|_{L^2(\Omega)} = (u, u)^{1/2},$$

$$((u, v)) = (u, v) + \sum_{i=1}^{2}\left(\frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i}\right), \quad \|u\| = ((u, u))^{1/2}.$$

2.1. Space decompositions. Since $V_{d-1} \subset V_d$, it is useful to define a supplement W_{d-1} of V_{d-1} into V_d so that

$$V_d = V_{d-1} \oplus W_d. \tag{2.2}$$

For that purpose we define W_d as the subspace of V_d consisting of the functions $z \in V_d$ that vanish at the coarse grid points, i.e.,

$$z(M) = 0 \quad \forall M \in \mathscr{U}_{d-1}. \tag{2.3}$$

It is clear that $V_{d-1} \cap W_{d-1}$ is reduced to $\{0\}$. Furthermore, since the functions of V_d are uniquely defined by their values at the nodal points on $\mathscr{U}_d$, every function $u \in V_d$ can be uniquely written as the sum

$$u = y + z, \quad y \in V_{d-1}, \quad z \in W_d, \tag{2.4}$$

so that

$$\begin{aligned} y(M) &= u(M) \quad \forall M \in \mathscr{U}_{d-1}, \\ z(P) &= u(P) - y(P) \quad \forall P \in \mathscr{U}_d \backslash \mathscr{U}_{d-1}. \end{aligned} \tag{2.5}$$

The rectangles $K \in \mathscr{T}_d$ are obtained by dividing the rectangles of $\mathscr{T}_{d-1}$ into four equal rectangles. Consider a rectangle $K \in \mathscr{T}_{d-1}$ and the corresponding rectangles of $\mathscr{T}_d$ denoted by K_1, K_2, K_3, K_4 as in Fig. 2.1. Since y is linear along the edges of K, we see with (2.5) that

$$\begin{aligned} z(P_1) &= u(P_1) - \tfrac{1}{2}(u(M_1) + u(M_2)), \\ z(P_2) &= u(P_2) - \tfrac{1}{2}(u(M_1) + u(M_3)), \\ z(P_3) &= u(P_3) - \tfrac{1}{4}(u(M_1) + u(M_2) + u(M_3) + u(M_4)). \end{aligned} \tag{2.6}$$

Hence the nodal values of z at the points $P_i \in \mathscr{U}_d \backslash \mathscr{U}_{d-1}$ are the incremental values of u as defined in § 1.

Of course, we can reiterate the decomposition (2.2),

$$V_{d-1} = V_{d-2} \oplus W_{d-2}, \ldots,$$

and finally,

$$V_d = V_0 \oplus W_1 \oplus W_2 \oplus \cdots \oplus W_d. \tag{2.7}$$

2.2. Interpolation. The linear interpolation operator $r_d = r_{h_d}$ associates to any continuous function on $\bar{\Omega}$, $u \in \mathscr{C}(\bar{\Omega})$, the function $r_d u \in V_d$ defined by its nodal values

$$r_d u(M) = u(M) \quad \forall M \in \mathscr{U}_d. \tag{2.8}$$

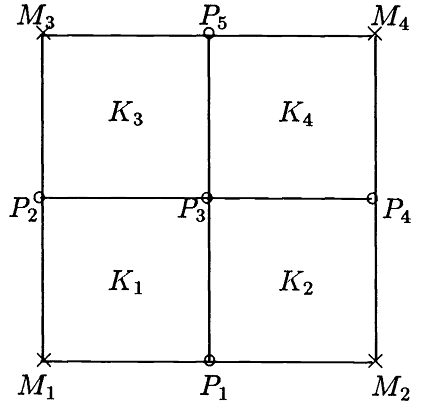


FIG. 2.1. *A rectangle K of $\mathcal{T}_{d-1}$ divided into four rectangles of $\mathcal{T}_d$.*

Similar interpolation operators $r_l = r_{h_l}$, $l = 0, \dots, d-1$, can be defined and if $u \in V_d$, $r_d u = u$, and the obvious formula

$$u = r_d u = r_0 u + \sum_{l=1}^{d} (r_l u - r_{l-1} u) \tag{2.9}$$

yields the decomposition of u corresponding to (2.7) as $r_0 u \in V_0$ and $r_l u - r_{l-1} u \in W_l$.

2.3. Other norms. If $u \in V_d$, $z = r_d u - r_{d-1} u$ vanishes at the nodes $M \in \mathcal{U}_{d-1}$ and takes the incremental values as in (2.6) at the nodes $P \in \mathcal{U}_d \backslash \mathcal{U}_{d-1}$; similar relations hold at the other levels of discretization. Hence the $(r_l u - r_{l-1} u)(x)$, $x \in \mathcal{U}_l \backslash \mathcal{U}_{l-1}$, $l = 1, \dots, d$, are the incremental values of u at the different levels. We then endow V_d with a semi-norm and a norm corresponding to the incremental values. If $u \in V_d$, we set

$$[u]_{hd}^2 = [u]_d^2 = \sum_{l=1}^{d} \sum_{x \in \mathcal{U}_l \backslash \mathcal{U}_{l-1}} |(r_l u - r_{l-1} u)(x)|^2, \tag{2.10}$$

$$[\![u]\!]_{h_d}^2 = [\![u]\!]_d^2 = [u]_d^2 + \sum_{x \in \mathcal{U}_0} |r_0 u(x)|^2. \tag{2.11}$$

To a function u of V_d, we associate the step function $\tilde{u}$ such that

$$\tilde{u}(x) = u(j_1 h_{1,d}, j_2 h_{2,d}) \quad \forall x \in K_{j,d},$$

$j = (j_1, j_2)$, $0 \leqq j_i \leqq N \cdot 2^d$. We then define the discrete H^1 scalar product and norm

$$\begin{aligned} ((u, v))_d &= ((u, v))_{h_d} \\ &= \int_\Omega \tilde{u}(x) \tilde{v}(x)\, dx + \sum_{i=1}^{2} \int_\Omega \nabla_{i,h_d} \tilde{u}(x) \nabla_{i,h_d} \tilde{v}(x)\, dx, \\ \|u\|_d &= \{((u, u))_d\}^{1/2} \quad \forall u, v \in V_d. \end{aligned}$$

Here ∇_{1,h_d}, ∇_{2,h_d} are finite difference operators

$$\nabla_{1,h_d} \varphi(x) = \frac{1}{h_{1,d}} \{\varphi(x_1 + h_{1,d}, x_2) - \varphi(x_1, x_2)\},$$

$$\nabla_{2,h_d} \varphi(x) = \frac{1}{h_{2,d}} \{\varphi(x_1, x_2 + h_{2,d}) - \varphi(x_1, x_2)\}.$$

The comparison of the norm $\|\cdot\|_d$ to the norm $\|\cdot\|$, uniform with respect to h_d, is the purpose of Lemmas 4.1 and 4.2.

At this point we can state the following equivalence norms theorem, which will be proved in §§ 3 and 4.

THEOREM 2.1. (i) *There exist two constants c_1 and c_2 that depend only on $\theta = h_{1,0}/h_{2,0}$ such that, for every $u \in V_d$,*

$$\tag{2.12} \begin{aligned} \frac{c_1}{(d+1)}\{|\nabla r_0 u|^2_{L^2(\Omega)} + [u]^2_d\}^{1/2} &\leqq |\nabla u|_{L^2(\Omega)} \\ &\leqq c_2\{|\nabla r_0 u|^2_{L^2(\Omega)} + [u]^2_d\}^{1/2}, \end{aligned}$$

where ∇ is the gradient operator.

(ii) *There exist two constants c_3 and c_4 that depend only on $\theta = h_{1,0}/h_{2,0}$, such that for every $u \in V_d$*

$$\tag{2.13} \begin{aligned} &\frac{c_3}{(d+1)}\{|\nabla r_0 u|^2_{L^2(\Omega)} + |\Omega|^{-1}|r_0 u|^2_{L^2(\Omega)} + [u]^2_d\}^{1/2} \\ &\leqq \{|\nabla u|^2_{L^2(\Omega)} + |\Omega|^{-1}|u|^2_{L^2(\Omega)}\}^{1/2} \\ &\leqq c_4\{|\nabla r_0 u|^2_{L^2(\Omega)} + |\Omega|^{-1}|r_0 u|^2_{L^2(\Omega)} + [u]^2_d\}^{1/2}, \end{aligned}$$

where $|\Omega|$ denotes the area of Ω.

Remark 2.1. Other inequalities can be obviously obtained for $u \in V_d \cap H^1_0(\Omega)$ by combining (2.12) or (2.13) with the Poincaré inequality.

3. Comparisons of norms. We start with an inequality concerning the L^∞ norm on V_d. Inequalities of this sort are due to [Th], [W], and [Y1], and in the case of spectral spaces, to [FMT] and [FMT2].

LEMMA 3.1. *There exist three positive constants c'_1, c'_2, and c'_3, depending only on $\theta = h_{1,0}/h_{2,0}$, such that*

$$\tag{3.1} |u|_{L^\infty(\Omega)} \leqq c'_1\left(\log\frac{|\Omega|}{h_{1,d}h_{2,d}}\right)^{1/2}|\nabla u|_{L^2(\Omega)} \quad \forall u \in V_d \cap H^1_0(\Omega),$$

and for every $u \in V_d$,

$$\tag{3.2} |u - \bar{u}|_{L^\infty(\Omega)} \leqq c'_2\left(\log\frac{|\Omega|}{h_{1,d}h_{2,d}}\right)^{1/2}|\nabla u|_{L^2(\Omega)},$$

$$\tag{3.3} |u|_{L^\infty(\Omega)} \leqq c'_3\left(\log\frac{|\Omega|}{h_{1,d}h_{2,d}}\right)^{1/2}\{|\nabla u|_{L^2(\Omega)} + |\Omega|^{-1/2}|u|_{L^2(\Omega)}\},$$

where $\bar{u}$ is the average of u on Ω and $|\Omega|$ is the area of Ω.

Proof. For (3.1) we start from the inequality

$$\tag{3.4} |u|_{L^p(\Omega)} \leqq cp^{1/2}|\Omega|^{1/p}|\nabla u|_{L^2(\Omega)} \quad \forall u \in H^1_0(\Omega),$$

which is valid for every p, $2 \leqq p < \infty$, with a constant c depending only on θ; $|\Omega|$ is the volume (area) of Ω. This inequality follows from [GTr], where according to [GTr, Lemma 7.14] and [GTr, Lemma 7.12, eq. (7.34)],

$$|u(x)| \leqq |g(x)| \quad \text{a.e.}, \qquad g(x) = \frac{1}{2\pi}\int_\Omega \frac{|\nabla u(\xi)|}{|x-\xi|}\,d\xi,$$

$$|g|_{L^p(\Omega)} \leqq cp^{1/2}|\Omega|^{1/p}|\nabla u|_{L^2(\Omega)}.$$

For u in $H^1(\Omega)$ we have the inequality

$$(3.5) \qquad |u - \bar{u}|_{L^p(\Omega)} \leqq c'p^{1/2}|\Omega|^{1/p}|\nabla u|_{L^2(\Omega)},$$

which follows similarly from [GTr, Lemmas 7.16 and 7.12, eq. (7.34)]; in (3.4) c is an absolute constant, while in (3.5) c' depends on θ. We also have, for every u in $H^1(\Omega)$,

$$(3.6) \qquad \begin{aligned} |u|_{L^p(\Omega)} &\leqq |u-\bar{u}|_{L^p(\Omega)} + |\bar{u}|_{L^p(\Omega)} \\ &\leqq c'p^{1/2}|\Omega|^{1/p}|\nabla u|_{L^2(\Omega)} + |\Omega|^{1/p-1/2}|u|_{L^2(\Omega)}. \end{aligned}$$

Now by the inverse-type inequalities for finite elements [C], [Th], there exists a constant c'' depending only on θ such that

$$(3.7) \qquad |u|_{L^\infty(\Omega)} \leqq c''(h_{1,d}h_{2,d})^{-1/p}|u|_{L^p(\Omega)} \quad \forall u \in V_d;$$

this inequality is obtained by mapping the element on which u achieves its maximum onto a reference element, e.g., Ω itself.

Finally, we combine (3.4), (3.5), or (3.6) with (3.7) and choose $p = \log|\Omega|/h_{1,d}h_{2,d}$; (3.1), (3.2), and (3.3) follow promptly. □

We then prove some technical lemmas.

LEMMA 3.2. *There exist two constants c_4' and c_5' depending only on θ, such that, for every $u \in V_d$ and every $0 \leqq l \leqq d$,*

$$(3.8) \qquad |\nabla r_l u|_{L^2(\Omega)} \leqq c_4'(d-l)^{1/2}|\nabla u|_{L^2(\Omega)},$$

$$(3.9) \qquad |r_l u|_{L^2(\Omega)} \leqq c_5'(d-l)^{1/2}\{h_{1,l}h_{2,l}|\nabla u|^2_{L^2(\Omega)} + |u|^2_{L^2(\Omega)}\}^{1/2}.$$

Proof. We first observe that Lemma 3.1 is valid if we replace Ω by a rectangle $K \in \mathscr{T}_l$. In this case $|\Omega|$ is replaced by $h_{1,l}h_{2,l}$ and

$$\frac{|\Omega|}{h_{1,d}h_{2,d}} = \frac{h_{1,l}h_{2,l}}{h_{1,d}h_{2,d}} = 4^{d-l}.$$

Hence (3.2) becomes

$$(3.10) \qquad |u - \bar{u}_K|_{L^\infty(K)} \leqq c(d-l)^{1/2}|\nabla u|_{L^2(K)},$$

where $\bar{u}_K$ is the average of u on K.

On the other hand, we have the inequality

$$(3.11) \qquad |\nabla r_l u|_{L^2(K)} \leqq c'|r_l u - (\overline{r_l u})_K|_{L^\infty(K)}, \qquad K \in \mathscr{T}_l,$$

which is valid with a constant c' depending only on θ since (3.11) is invariant by homothety. In fact, in (3.11), we can replace the average of $r_l u$ on K by any real number α:

$$(3.12) \qquad |\nabla r_l u|_{L^2(K)} \leqq c' \operatorname*{Inf}_{\alpha \in \mathbb{R}} |r_l u - \alpha|_{L^\infty(K)}, \qquad K \in \mathscr{T}_l.$$

We then choose $\alpha = \bar{u}_K$ and observe that $r_l u$ agrees with u at the vertices of K. Therefore,

$$(3.13) \qquad |r_l u - \bar{u}_K|_{L^\infty(K)} \leqq |u - \bar{u}_K|_{L^\infty(K)},$$

and comparing (3.10), (3.12), and (3.13) we find

$$|\nabla r_l u|^2_{L^2(K)} \leqq c''(d-l)|\nabla u|^2_{L^2(K)} \quad \forall K \in \mathscr{T}_l;$$

(3.8) follows by summation for $K \in \mathscr{T}_l$.

The proof of (3.9) is similar. First replacing Ω by $K \in \mathscr{T}_l$, (3.3) becomes

$$|u|_{L^\infty(K)} \leqq c(d-l)^{1/2}\{|\nabla u|_{L^2(K)} + (h_{1,l}h_{2,l})^{-1/2}|u|_{L^2(K)}\}. \tag{3.14}$$

Then we write

$$\begin{aligned}|r_l u|^2_{L^2(K)} &\leqq (h_{1,l}h_{2,l})|u|^2_{L^\infty(K)}\\ &\leqq (\text{with } (3.14))\\ &\leqq c'(d-l)\{h_{1,l}h_{2,l}|\nabla u|^2_{L^2(K)} + |u|^2_{L^2(K)}\}.\end{aligned}$$

Adding these relations for $K \in \mathscr{T}_l$, we obtain (3.9).

LEMMA 3.3. *There exist two constants c_6' and c_7' depending only on $\theta = h_{1,0}/h_{2,0}$ such that*

$$c_6'[u]_d^2 \leqq \sum_{k=1}^{d} |\nabla(r_k u - r_{k-1}u)|^2_{L^2(\Omega)} \leqq c_7'[u]_d^2 \quad \forall u \in V_d. \tag{3.15}$$

Proof. For a fixed k, $1 \leqq k \leqq d$, there exist two constants c and c' depending only on θ and k such that, for every $v \in W_k$,

$$c \sum_{x \in \mathscr{U}_k \backslash \mathscr{U}_{k-1}} |v(x)|^2 \leqq |\nabla v|^2_{L^2(\Omega)} \leqq c' \sum_{x \in \mathscr{U}_k \backslash \mathscr{U}_{k-1}} |v(x)|^2. \tag{3.16}$$

Indeed, let K be a rectangle of $\mathscr{T}_k$, as in Fig. 2.1 (with $d-1$ replaced by k). Then there exist c and c' such that

$$c \sum_{i=1}^{5} |\varphi(P_i)|^2 \leqq |\nabla\varphi|^2_{L^2(K)} \leqq c' \sum_{i=1}^{5} |\varphi(P_i)|^2 \tag{3.17}$$

for every continuous function φ which is Q_1 on $K_1, \dots, K_4$ and vanishes at $M_1, \dots, M_4$.

Indeed, $|\nabla\varphi|_{L^2(K)}$ is a norm on the set of such functions φ, as is the expression $\{\sum_{i=1}^5 |\varphi(P_i)|^2\}^{1/2}$. Furthermore, since (3.17) is invariant by homothety, we observe by mapping K onto a reference rectangle that c and c' are independent of h_k and depend only on $\theta = h_{1,k}/h_{2,k}$.

We then write (3.17) with $\varphi = v$ and add these relations for $K \in \mathscr{T}_k$. Since each node P_i belongs to at most two different rectangles $K \in \mathscr{T}_k$, we obtain (3.16). Finally, we write (3.16) with $v = r_k u - r_{k-1}u \in W_k$, and by summation of these relations for $k = 1, \dots, d$, we find (3.15). □

We can now prove the first inequality in (2.12).

LEMMA 3.4. *The first inequalities in* (2.12) *and* (2.13) *are valid.*

Proof. Due to (3.15),

$$\begin{aligned}|\nabla r_0 u|^2_{L^2(\Omega)} + [u]_d^2 &\leqq c\left\{|\nabla r_0 u|^2_{L^2(\Omega)} + \sum_{k=1}^{d} |\nabla(r_k u - r_{k-1}u)|^2_{L^2(\Omega)}\right\}\\ &\leqq c\left\{|\nabla r_0 u|^2_{L^2(\Omega)} + 2\sum_{k=1}^{d} |\nabla r_k u|^2_{L^2(\Omega)} + |\nabla r_{k-1}u|^2_{L^2(\Omega)}\right\}\\ &\leqq 4c\sum_{k=0}^{d} |\nabla r_k u|^2_{L^2(\Omega)}\\ &\leqq (\text{with } (3.8))\\ &\leqq c'\sum_{k=0}^{d} (d-k)|\nabla u|^2_{L^2(\Omega)}\\ &\leqq c'd(d+1)|\nabla u|^2_{L^2(\Omega)}.\end{aligned} \tag{3.18}$$

For the first inequality (2.13), we infer from (3.9), written with $l = 0$,

$$(3.19) \qquad |\Omega|^{-1}|r_0 u|^2_{L^2(\Omega)} \leqq (c_5')^2 d \left\{ \frac{h_{1,0}h_{2,0}}{|\Omega|} |\nabla u|^2_{L^2(\Omega)} + |\Omega|^{-1}|u|^2_{L^2(\Omega)} \right\}.$$

Since $h_{1,0}h_{2,0} \leqq |\Omega|$, we obtain the first inequality in (2.13) by adding (3.18) and (3.19).

4. Proof of Theorem 2.1. The proof of the second inequality in (2.12) and (2.13) follows from Lemmas 4.4 and 4.5. The first three lemmas provide a comparison of finite element and finite difference norms; more precisely, Lemmas 4.1 and 4.2 provide a comparison of the norms in $L^2(\Omega)$ of ∇u_k and $\nabla_h \tilde{u}_k$, $\tilde{u}_k$ defined in § 2. Lemma 4.3 gives a comparison of the L^2 norms of u_h and $\tilde{u}_h$. Then Lemma 4.5 provides an enhanced Cauchy–Schwarz inequality.

LEMMA 4.1. *For every* $u \in V_d$,

$$(4.1) \qquad \tfrac{1}{2}|\nabla_{h_d}\tilde{u}|_{L^2(\Omega)} \leqq |\nabla u|_{L^2(\Omega)}.$$

Proof. To prove (4.1), it suffices to show that

$$(4.2) \qquad \frac{1}{4}\int_K |\nabla_{h_d}\tilde{u}|^2\,dx \leqq \int_K |\nabla u|^2\,dx,$$

where K is any of the rectangles of $\mathcal{T}_d$.

Assume for simplicity that K is the rectangle $(0, \alpha_1) \times (0, \alpha_2)$ $(\alpha_1 = h_{1,d}, \alpha_2 = h_{2,d})$, with vertices A_1, A_2, A_3, and A_4, as shown in Fig. 4.1.

We then immediately see that $\nabla_{h_d}\tilde{u}$ is a constant on K and

$$\nabla_{h_d}\tilde{u} = \begin{pmatrix} \dfrac{1}{\alpha_1}\{u(A_2) - u(A_1)\} \\ \dfrac{1}{\alpha_2}\{u(A_4) - u(A_1)\} \end{pmatrix},$$

$$(4.3) \qquad \int_K |\nabla_{h_d}\tilde{u}|^2\,dx = \alpha_1\alpha_2\left\{\frac{1}{\alpha_1^2}(u(A_2) - u(A_1))^2 + \frac{1}{\alpha_2^2}(u(A_4) - u(A_1))^2\right\}.$$

On the other hand, a lengthy but straightforward calculation shows that, on K,

$$(4.4) \qquad \nabla u = \begin{pmatrix} \left(1 - \dfrac{x_2}{\alpha_2}\right)\dfrac{1}{\alpha_1}(u(A_2) - u(A_1)) + \dfrac{x_2}{\alpha_2}\dfrac{1}{\alpha_1}(u(A_3) - u(A_4)) \\ \left(1 - \dfrac{x_1}{\alpha_1}\right)\dfrac{1}{\alpha_2}(u(A_4) - u(A_1)) + \dfrac{x_1}{\alpha_1}\dfrac{1}{\alpha_2}(u(A_3) - u(A_2)) \end{pmatrix}$$

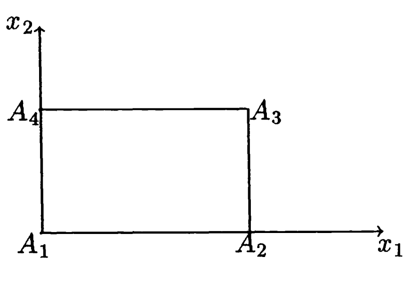


FIG. 4.1. *A rectangle K of* $\mathcal{T}_d$.

and

$$\frac{1}{\alpha_1\alpha_2}\int_K |\nabla u|^2\,dx = \sigma_1^2+\sigma_2^2+(\sigma_1(\tau_1-\sigma_1)+\sigma_2(\tau_2-\sigma_2)) + \frac{1}{3}((\tau_1-\sigma_1)^2+(\tau_2-\sigma_2)^2), \tag{4.5}$$

where

$$\sigma_1=\frac{1}{\alpha_1}(u(A_2)-u(A_1)),\qquad \sigma_2=\frac{1}{\alpha_2}(u(A_4)-u(A_1)),$$
$$\tau_1=\frac{1}{\alpha_1}(u(A_3)-u(A_4)),\qquad \tau_2=\frac{1}{\alpha_2}(u(A_3)-u(A_2)). \tag{4.6}$$

Hence with Cauchy-Schwarz inequality,

$$\frac{1}{\alpha_1\alpha_2}\int_K |\nabla u|^2\,dx \geqq \frac{1}{4}\{(\sigma_1^2+\sigma_2^2)+(\tau_1-\sigma_1)^2+(\tau_2-\sigma_2)^2\},$$

and (4.2) follows. □

For a more precise inequality and in order to establish the inverse inequality of (4.1), we must introduce the extended norms

$$|\nabla_{1,h}^e\tilde{u}|_{L^2(\Omega)},\qquad |\nabla_{2,h}^e\tilde{u}|_{L^2(\Omega)}.$$

For this purpose we observe that even if u (and $\tilde{u}$) are defined on $\Omega=(0,a_1)\times(0,a_2)$, the finite difference quotient $\nabla_{1,h}\tilde{u}$ is defined on the extended rectangle $(0,a_1)\times(0,a_2+h_2)$, and $\nabla_{2,h}\tilde{u}$ is defined on the extended rectangle $(0,a_1+h_1)\times(0,a_2)$. For the sake of simplicity, we agree to write

$$|\nabla_{1,h}^e\tilde{u}|_{L^2(\Omega)}=\left\{\int_{\Omega\cup(\Omega+h_2e_2)}|\nabla_{1,h}\tilde{u}(x)|^2\,dx\right\}^{1/2},$$

$$|\nabla_{2,h}^e\tilde{u}|_{L^2(\Omega)}=\left\{\int_{\Omega\cup(\Omega+h_1e_1)}|\nabla_{2,h}\tilde{u}(x)|^2\,dx\right\}^{1/2},$$

$$|\nabla_h^e\tilde{u}|_{L^2(\Omega)}=\{|\nabla_{1,h}^e\tilde{u}|_{L^2(\Omega)}^2+|\nabla_{2,h}^e\tilde{u}|_{L^2(\Omega)}^2\}^{1/2},$$

$$e_1=(1,0),\qquad e_2=(0,1).$$

Then we can state the following lemma.

LEMMA 4.2. *For every $u\in V_d$,*

$$\frac{1}{\sqrt{6}}|\nabla_{h_d}^e\tilde{u}|_{L^2(\Omega)}\leqq|\nabla u|_{L^2(\Omega)}\leqq\sqrt{6}|\nabla_{h_d}^e\tilde{u}|_{L^2(\Omega)}. \tag{4.7}$$

Proof. As in Lemma 4.1, let K be a rectangle of $\mathscr{T}_d$. Then (4.3) and (4.5) are valid. Using the Cauchy-Schwarz inequality, we can bound the right-hand side of (4.5) as follows:

$$\frac{1}{\alpha_1\alpha_2}\int_K|\nabla u|^2\,dx\leqq 2(\sigma_1^2+\sigma_2^2)+\left(\frac{1}{3}+\frac{1}{4}\right)((\tau_1-\sigma_1)^2+(\tau_2-\sigma_2)^2)$$
$$\leqq 4(\sigma_1^2+\sigma_2^2)+2(\tau_1^2+\tau_2^2),$$

$$\int_K|\nabla u|^2\,dx\leqq 4\int_K|\nabla_{h_d}\tilde{u}|^2\,dx+2\int_{K+h_{2,d}e_2}|\nabla_{1,h_d}\tilde{u}|^2\,dx + 2\int_{K+h_{1,d}e_1}|\nabla_{2,h_d}\tilde{u}|^2\,dx. \tag{4.8}$$

By adding inequalities (4.8) for all K's in $\mathcal{T}_d$, we find

$$\int_K |\nabla u|^2\, dx \leqq 6 \int_{\Omega\cup(\Omega+h_{2,d}e_2)} |\nabla_{1,h_d}\tilde{u}|^2\, dx + 6 \int_{\Omega\cup(\Omega+h_{1,d}e_1)} |\nabla_{2,h_d}\tilde{u}|^2\, dx,$$

and this is precisely the second inequality (4.7).

For the first inequality (4.7), we must treat the derivatives $D_i u = \partial u/\partial x_i$ separately. As for (4.5) we see that

$$\begin{aligned}\frac{1}{\alpha_1\alpha_2}\int_K |D_1 u|^2\, dx &= \sigma_1^2 + (\sigma_1(\tau_1 - \sigma_1)) + \frac{1}{3}(\tau_1 - \sigma_1)^2\\ &= \frac{1}{3}(\tau_1^2 + \sigma_1^2 + \tau_1\sigma_1)\\ &\geqq \frac{1}{6}(\sigma_1^2 + \tau_1^2),\end{aligned}$$

$$\int_K |D_1 u|^2\, dx \geqq \frac{1}{6}\int_K |\nabla_{1,h_d}\tilde{u}|^2\, dx + \frac{1}{6}\int_{K+h_{2,d}e_2} |\nabla_{1,h_d}\tilde{u}|^2\, dx.$$

Similarly, in the direction x_2,

$$\int_K |D_2 u|^2\, dx \geqq \frac{1}{6}\int_K |\nabla_{2,h_d}\tilde{u}|^2\, dx + \frac{1}{6}\int_{K+h_{1,d}e_1} |\nabla_{2,h_d}\tilde{u}|^2\, dx.$$

By adding all these inequalities for $K \in \mathcal{T}_d$, we obtain

$$\int_\Omega |\nabla u|^2\, dx \geqq \frac{1}{6}\int_{\Omega\cup(\Omega+h_{2,d}e_2)} |\nabla_{1,h_d}\tilde{u}|^2\, dx + \frac{1}{6}\int_{\Omega\cup(\Omega+h_{1,d}e_1)} |\nabla_{2,h_d}\tilde{u}|^2\, dx,$$

and this is the first inequality (4.7).

Remark 4.1. Let us observe that for functions vanishing on $\partial\Omega$, as is the case for a Dirichlet problem, the extended domains are not needed and

$$|\nabla^e_{i,h}\tilde{u}|_{L^2(\Omega)} = |\nabla_{i,h}\tilde{u}|_{L^2(\Omega)}, \qquad i = 1, 2,$$
$$|\nabla^e_h\tilde{u}|_{L^2(\Omega)} = |\nabla_h\tilde{u}|_{L^2(\Omega)}.$$

In particular, the following simplifications arise. The left inequality of (4.7) is not needed (and is replaced by (4.1)), and for the second inequality in (4.7), we can replace $\nabla^e_{h_d}\tilde{u}$ by $\nabla_{h_d}\tilde{u}$. □

We now prove the following result concerning the L^2 norms.

LEMMA 4.3. *For every $u \in V_d$, there exist two constants c_8' and c_9' depending only on $\theta = h_{1,0}/h_{2,0}$, such that*

$$c_8'\left(\int_{\Omega^*_{h_d}} |\tilde{u}|^2\, dx\right)^{1/2} \leqq |u|_{L^2(\Omega)} \leqq c_9'\left(\int_{\Omega^*_{h_d}} |\tilde{u}|^2\, dx\right)^{1/2}, \tag{4.9}$$

where $\Omega^*_{h_d}$ is an extension of Ω; $\Omega^*_{h_d} = (0, a_1 + h_{1,d}) \times (0, a_2 + h_{2,d})$.

Proof. Let K be a rectangle of $\mathcal{T}_d$ as in Fig. 4.1; $|u|_{L^2(K)}$ and $\{h_{1,d}h_{2,d} \times \sum_{i=1}^4 |u(A_i)|^2\}^{1/2}$ are two equivalent norms on the four-dimensional space consisting of the Q_1 functions on K. There exist two constants c and c' depending only on θ and d, such that

$$c\left\{h_{1,d}h_{2,d}\sum_{i=1}^{4} |u(A_i)|^2\right\}^{1/2} \leqq |u|_{L^2(K)} \leqq c'\left\{h_{1,d}h_{2,d}\sum_{i=1}^{4} |u(A_i)|^2\right\}^{1/2}. \tag{4.10}$$

Furthermore, since formula (4.10) is invariant under homothety, we observe by mapping K onto a reference rectangle that c, c' are independent of h_d and depend only on $\theta = h_{1,0}/h_{2,0}$. Since

$$\int_{K^*_{h_d}} |\tilde{u}|^2\,dx = h_{1,d}h_{2,d} \sum_{i=1}^{4} |u(A_i)|^2$$

and each node A_i belongs to at most four different $K^*_{h_d}$, summation for all $K \in \mathscr{T}_d$ gives

$$c\int_{\Omega^*_{h_d}} |\tilde{u}|^2\,dx \leqq |u|_{L^2(\Omega)} \leqq 4c' \int_{\Omega^*_{h_d}} |\tilde{u}|^2\,dx.$$

LEMMA 4.4. *There exists a positive constant c'_{10} depending only on $\theta = h_{1,0}/h_{2,0}$ such that for every $u \in V_d$*

$$(4.11) \qquad |u|_{L^2(\Omega)} \leqq c'_{10}\{|r_0u|^2_{L^2(\Omega)} + |\Omega|[u]^2_d\}^{1/2}.$$

Proof. For every u in V_d, we write, as in (2.9),

$$u = r_du = r_0u + \sum_{l=1}^{d} \{r_lu - r_{l-1}u\},$$

where $r_lu - r_{l-1}u \in W_l \subset V_l$. Therefore,

$$(4.12) \qquad |u|_{L^2(\Omega)} \leqq |r_0u|_{L^2(\Omega)} + \sum_{l=1}^{d} |r_lu - r_{l-1}u|_{L^2(\Omega)}.$$

Now if $v \in V_l$,

$$|v|^2_{L^2(\Omega)} = \int_\Omega |v(x)|^2\,dx = \sum_{K\in\mathscr{T}_1} \int_K |v(x)|^2\,dx.$$

By using (4.10) with l replacing d and v replacing u,

$$\begin{aligned} |v|^2_{L^2(K)} &\leqq (c')^2 h_{1,l}h_{2,l} \sum_{i=1}^{4} |v(A_i)|^2 \leqq \frac{(c')^2}{2^{2l}} h_{1,0}h_{2,0} \sum_{i=1}^{4} |v(A_i)|^2 \\ (4.13) \qquad &\leqq \frac{(c')^2}{2^{2l}} h_{1,0}h_{2,0} \sum_{i=1}^{4} |v(A_i)|^2 \\ &\leqq \frac{(c')^2}{2^{2l}} |\Omega| \sum_{i=1}^{4} |v(A_i)|^2. \end{aligned}$$

We add inequalities (4.13) for $K \in \mathscr{T}_l$. Since each vertex A_i belongs to at most four rectangles K, we find

$$|v|^2_{L^2(\Omega)} \leqq \frac{4(c')^2}{2^{2l}} |\Omega| \sum_{x\in\mathscr{U}_l} |v(x)|^2 \quad \forall v \in V_l.$$

If v belongs to W_l, then $v(x) = 0$ for all $x \in \mathscr{U}_{l-1}$ and the previous inequality yields

$$(4.14) \qquad |v|^2_{L^2(\Omega)} \leqq \frac{4(c')^2}{2^{2l}} |\Omega|^{1/2} \sum_{x\in\mathscr{U}_l\setminus\mathscr{U}_{l-1}} |v(x)|^2 \quad \forall v \in V_{l-1}.$$

We now replace v by $r_l u - r_{l-1}u$ and use this bound in (4.12):

$$|u|_{L^2(\Omega)} \leqq |r_0 u|_{L^2(\Omega)} + c''|\Omega|^{1/2}\left\{\sum_{l=1}^{d} \frac{1}{2^l}\left\{\sum_{x\in\mathcal{U}_l\setminus\mathcal{U}_{l-1}} |(r_l u - r_{l-1}u)(x)|^2\right\}^{1/2}\right\}$$

$$\leqq |r_0 u|_{L^2(\Omega)} + c''|\Omega|^{1/2}\left\{\sum_{l=1}^{d} 4^{-l}\right\}^{1/2}\left\{\sum_{l=1}^{d}\sum_{x\in\mathcal{U}_l\setminus\mathcal{U}_{l-1}} |(r_l u - r_{l-1}u)(x)|^2\right\}^{1/2}$$

$$\leqq |r_0 u|_{L^2(\Omega)} + c''|\Omega|^{1/2}[u]_d.$$

Hence (4.11) holds. □

We now prove an enhanced Cauchy–Schwarz inequality for Q_1 finite element spaces; similar inequalities appear in [B], [MT], or [Y1].

LEMMA 4.5. *There exist two positive constants* c'_{11} *and* c'_{12} *depending only on* θ, $c'_{11} < 1$, *such that for every* u *in* V_k *and* v *in* V_l,

$$\int_\Omega \nabla u \nabla v \, dx \leqq \min\,(c'_{11}, c'_{12}2^{-|l-k|/2})|\nabla u|_{L^2(\Omega)}|\nabla v|_{L^2(\Omega)}. \tag{4.15}$$

Proof. Without loss of generality, we can assume that $k = 0$, $l \geqq 1$. The inequality (4.15) with $c'_{11} < 1$ follows readily from [MT, Lemma 3.2], by simply observing that if $u \in V_k$, then $u \in V_{l-1}$.

In order to prove the second inequality in (4.15), we first start with the case where Ω is replaced by K, K being a rectangle of $\mathcal{T}_0$ divided into 2^{2l} rectangles of $\mathcal{T}_l$. For simplicity we assume that K is the rectangle $(0, h_1) \times (0, h_2)$. The value of ∇u on K and the value of the integral

$$\int_K |\nabla u|^2 \, dx = |\nabla u|^2_{L^2(K)}$$

are given in (4.4) and (4.5) with α_1, α_2 replaced by $h_{1,0}$ and $h_{2,0}$.

The function v is continuous and is Q_1 on the $\mathcal{T}_l$ subrectangles of K. Let v_0 be the function of the same type that agrees with v at the nodes of $\mathcal{U}_l$ in the interior of K; then $v_1 = v - v_0$ vanishes on all these nodes and agrees with v on the nodes of $\mathcal{U}_l$ belonging to the boundary of K. Since u is Q_1 on K, it is a harmonic function and therefore, by integration by parts, since $v_0 = v$ on ∂K,

$$\int_\Omega \nabla u \nabla v_0 \, dx = 0;$$

thus

$$\int_\Omega \nabla u \nabla v \, dx = \int_\Omega \nabla u \nabla v_1 \, dx. \tag{4.16}$$

Now v_1 and ∇v_1 vanish except in the layer $\mathscr{E}$ of size $2^{-l}h_{1,0}$ or $2^{-l}h_{2,0}$ contiguous to ∂K. Hence

$$\int_\Omega \nabla u \nabla v \, dx = \int_{\mathscr{E}} \nabla u \nabla v_1 \, dx \leqq |\nabla u|_{L^2(\mathscr{E})}|\nabla v_1|_{L^2(\mathscr{E})}.$$

By mapping the $\mathscr{T}_l$ rectangles of $\mathscr{E}$ onto a reference rectangle we see easily that there exists a constant c' depending only on θ such that

$$\begin{aligned} |\nabla v_1|^2_{L^2(\mathscr{E})} &\leqq c' \sum_{x \in (\mathscr{U}_l \backslash \mathscr{U}_{l-1}) \cap \partial K} |v_1(x)|^2 \\ &= c' \sum_{x \in (\mathscr{U}_l \backslash \mathscr{U}_{l-1}) \cap \partial K} |v(x)|^2 \\ &\leqq c' \sum_{x \in (\mathscr{U}_l \backslash \mathscr{U}_{l-1}) \cap K} |v(x)|^2. \end{aligned}$$

Thanks to (3.16), the last quantity is majorized, with an appropriate constant $c'' = c''(\theta)$, by

$$c''|\nabla v|^2_{L^2(K)}.$$

The integral of $|\nabla u|^2$ on $\mathscr{E}$ is computed by straightforward computations; using (4.4), (4.5), and (4.6),

$$\begin{aligned} |\nabla u|^2 = \sigma_1^2 + \sigma_2^2 &+ 2\sigma_1 \frac{x_2}{\alpha_2}(\tau_1 - \sigma_1) + 2\sigma_2 \frac{x_1}{\alpha_1}(\tau_2 - \sigma_2) \\ &+ \frac{x_2}{\alpha_2^2}(\tau_1 - \sigma_1)^2 + \frac{x_1^2}{\alpha_1^2}(\tau_2 - \sigma_2)^2, \end{aligned}$$

$$\begin{aligned} \int_{\mathscr{E}} |\nabla u|_1^2\, dx = |\mathscr{E}|\{\sigma_1^2 + \sigma_2^2 + \sigma_1(\tau_1 - \sigma_1) + (\tau_2 - \sigma_2) \\ + r(\tau_1 - \sigma_1)^2 + r(\tau_2 - \sigma_2)^2\}, \end{aligned}$$

with $|\mathscr{E}|$ = area of $\mathscr{E} = 4h_{1,0}h_{2,0}2^{-l}(1 - 2^{-l})$ and

$$r = \frac{5 - 9\cdot 2^{-l} + 8\cdot 2^{-2l} - 4\cdot 2^{-3l}}{12(1 - 2^{-l})} \leqq \frac{3}{2}.$$

Hence

$$\frac{|\nabla u|^2_{L^2(\mathscr{E})}}{|\nabla u|^2_{L^2(K)}} \leqq 4\cdot 2^{-l}\delta,$$

where δ is the following ratio of positive definite quadratic forms (see (4.5)):

$$\delta = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_1(\tau_1 - \sigma_1) + \sigma_2(\tau_2 - \sigma_2) + \frac{3}{2}((\tau_1 - \sigma_1)^2 + (\tau_2 - \sigma_2)^2)}{\sigma_1^2 + \sigma_2^2 + (\sigma_1(\tau_1 - \sigma_1) + \sigma_2(\tau_2 - \sigma_2)) + \frac{1}{3}((\tau_1 - \sigma_1)^2 + (\tau_2 - \sigma_2)^2)}.$$

Since δ is clearly bounded from above by an absolute constant c''', we find

$$|\nabla u|^2_{L^2(\mathscr{E})} \leqq 4c'''2^{-l}|\nabla u|^2_{L^2(K)}.$$

Finally,

$$\int_K \nabla u \nabla v\, dx \leqq 2(c''c''')^{1/2}2^{-l/2}|\nabla u|_{L^2(K)}|\nabla v|_{L^2(K)}. \tag{4.17}$$

We then obtain (4.15) with $c'_{12} = 2(c'c''')^{1/2}$ by adding all these inequalities for $K \in \mathscr{T}_0$ and by using the Cauchy-Schwarz inequality for the sum in K in the right-hand side. □

We are now in position to complete the proof of Theorem 2.1.

LEMMA 4.6. *The second inequalities in* (2.12) *and* (2.13) *are valid.*

Proof. For $u \in V_d$ we infer from (2.9) that

$$u = r_d u = \sum_{j=0}^{d} v_j$$

with $v_0 = r_0 u$, $v_l = r_l u - r_{l-1} u$, and $1 \leqq l \leqq d$. Thus by Lemma 4.5,

$$\begin{aligned} |\nabla u|^2_{L^2(\Omega)} &= \left| \sum_{j=0}^{d} \nabla v_j \right|^2_{L^2(\Omega)} \\ &\leqq \sum_{j,k=0}^{d} (\nabla v_j, \nabla v_k) \\ &\leqq c'_{12} \sum_{j,k=0}^{d} 2^{-|j-k|/2} |\nabla v_j|_{L^2(\Omega)} |\nabla v_k|_{L^2(\Omega)}. \end{aligned}$$

The right-hand side of the last inequality can be written as $c'_{12} \sum_{j,k=0}^{d} \alpha_{jk} \eta_j \eta_k$ with $\alpha_{jk} = 2^{-|j-k|/2}$, $\eta_j = |\nabla v_j|_{L^2(\Omega)}$. Hence this is bounded by

$$c'_{12} \bar{\lambda} \sum_{j=0}^{d} \eta_j^2 = c'_{12} \bar{\lambda} \sum_{j=0}^{d} |\nabla v_j|^2_{L^2(\Omega)},$$

where $\bar{\lambda}$ is the largest eigenvalue of the matrix $\{\alpha_{jk}\}$. It is well known that $\bar{\lambda}$ is majorized by the largest row sum of the matrix $\{\alpha_{jk}\}$ and this row sum is majorized by

$$1 + 2 \sum_{j=1}^{\infty} 2^{-j/2} = \frac{\sqrt{2} + 1}{\sqrt{2} - 1}.$$

Using (3.15), we then find

$$|\nabla u|^2_{L^2(\Omega)} \leqq c |\nabla r_0 u|^2_{L^2(\Omega)} + c' [u]^2_d.$$

Hence (2.12) holds. The second inequality (2.13) is an immediate consequence of the second inequality (2.12) and Lemma 4.4.

5. IUs for boundary value problems. We now apply the previous results to the numerical solution of second-order boundary value problems by finite differences. We start with the simple case of the Dirichlet problem in the rectangle.

5.1. Dirichlet problem. We first consider the rectangle $\Omega = (0, a_1) \times (0, a_2)$ as before. The Dirichlet problem reads

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{5.1}$$

Let $h = (h_1, h_2)$ be the discretization mesh. Then the discretization of (5.1) by the usual five-points finite difference scheme leads to the classical linear system

$$A_h U_h = b_h. \tag{5.2}$$

Here A_h is the well-known symmetric positive definite matrix, U_h is the set of approximate nodal values of u, $u(ih_1, jh_2)$ properly ordered, and b_h is the set of values of f at points ih_1, jh_2, ordered in the same way.

Now assume that a multilevel discretization is used with meshes $h_0, \ldots, h_d$ as before, $h_l = (h_{1,l}, h_{2,l})$, $h_{i,l} = h_{i,0}/2^l$, $i = 1, 2$, $l = 0, \ldots, d$. We then have several systems (5.2),

$$A_{h_l} U_{h_l} = b_{h_l}, \qquad 0 \leqq l \leqq d, \tag{5.3}$$

but we are in fact interested in that corresponding to the finest mesh, $l = d$:

$$A_{h_d} U_{h_d} = b_{h_d}. \tag{5.4}$$

As we did before, and in order to make the notations simple, instead of (5.3), (5.4), we will write

$$A_l U_l = b_l, \qquad 0 \leqq l \leqq d. \tag{5.5}$$

At each level l, the set of incremental unknowns $\bar{U}$ consists of the following:

$$\bar{U}_l = \begin{bmatrix} Y^l \\ Z^l \end{bmatrix}, \qquad Z^l = \begin{bmatrix} Z_1 \\ \vdots \\ Z_l \end{bmatrix}.$$

- $Y^l = Y_0 =$ the set, properly ordered, of the approximate nodal values of u at the coarse grid;
- $Z_j =$ the set, properly ordered, of the incremental unknowns at the level j.

Hence, using the notations of § 2, if $u_h \in V_d$ is the approximate function, then $Y^l = Y_0$ consists of the values of $u_h(x)$, $x \in \mathcal{U}_0$, while Z_j consists of the values of $r_j u_h(x) - r_{j-1} u_h(x)$ for $x \in \mathcal{U}_j \backslash \mathcal{U}_{j-1}$.

Each system (5.5) plays the same role as (1.1) and we can pass from U_l to $\bar{U}_l$ by using a transfer matrix S_l. Hence we find, as in (1.2), (1.3),

$$U_l = S_l \bar{U}_l,$$

$$A_l S_l \bar{U}_l = b_l,$$

$$\bar{A}_l \bar{U}_l = \bar{b}_l, \qquad 0 \leqq l \leqq d, \tag{5.6}$$

with $\bar{A}_l = {}^t S_l A_l S_l$, $\bar{b}_l = {}^t S b_l$.

As we said above, we are interested in solving (5.6) when $l = d$ and we are interested in estimating the condition number of $\bar{A}_d$ and comparing it to the condition number of A_d. Since $\bar{A}_d$ is symmetric positive definite, its condition number is the ratio of its largest eigenvalue $\bar{\lambda}(\bar{A}_d)$ to its smallest eigenvalue $\underline{\lambda}(\bar{A}_d)$:

$$\kappa(\bar{A}_d) = \bar{\lambda}(\bar{A}_d) / \underline{\lambda}(\bar{A}_d). \tag{5.7}$$

As is well known, we can estimate $\underline{\lambda}$ and $\bar{\lambda}$ through some bounds of the associated quadratic form $\langle \bar{A}_d \bar{U}, \bar{U} \rangle$, where $\langle , \rangle$ denotes the usual Euclidean scalar product

$$\begin{aligned} \underline{\lambda}(\bar{A}_d) &= \operatorname*{Inf}_{U \neq 0} \frac{\langle \bar{A}_d U, U \rangle}{\langle U, U \rangle}, \\ \bar{\lambda}(\bar{A}_d) &= \operatorname*{Sup}_{U \neq 0} \frac{\langle \bar{A}_d U, U \rangle}{\langle U, U \rangle}. \end{aligned} \tag{5.8}$$

The bilinear form $\langle \bar{A}_d U, U^* \rangle$ arises naturally in the variational formulation of finite differences. More precisely (see [Ce] and [T1]),

$$\langle A_d U_d, U_d^* \rangle = \int_\Omega \nabla_{h_d} \tilde{u}_{h_d} \nabla_{h_d} \tilde{u}^*_{h_d} \, dx, \tag{5.9}$$

where $\tilde{u}_{h_d}$ and $\tilde{u}^*_{h_d}$ are the step functions associated with U_d and U^*_d (see § 2 and above). Therefore, if

$$U_d = S_d\bar{U}_d, \qquad U^*_d = S_d\bar{U}^*_d,$$

then

$$\langle \bar{A}_d\bar{U}_d, \bar{U}^*_d\rangle = \langle A_dU_d, U^*_d\rangle, \tag{5.10}$$

so that the right-hand side of (5.9) also provides the bilinear form associated with $\bar{A}_d$.

We readily infer from (2.12) (Theorem 2.1), Lemmas 4.1 and 4.2, and Remark 4.1 that

$$\begin{aligned}\langle \bar{A}_d\bar{U}_d, \bar{U}_d\rangle &= \int_\Omega |\nabla_{h_d}\tilde{u}_{h_d}|^2\,dx\\ &\geqq \frac{1}{6}\int_\Omega |\nabla u_{h_d}|^2\,dx\\ &\geqq \frac{c_1^2}{6(d+1)^2}\{|\nabla r_0u_{h_d}|^2_{L^2(\Omega)} + [u_{h_d}]^2_d\}\\ &\geqq \frac{c_1^2}{6(d+1)^2}\left\{\frac{1}{4}|\nabla_{h_0}(r_0\widetilde{u_{h_d}})|^2_{L^2(\Omega)} + [u_{h_d}]^2_d\right\}.\end{aligned} \tag{5.11}$$

We observe that

$$|\nabla_{h_0}(r_0\widetilde{u_{h_d}})|^2_{L^2(\Omega)} = \langle A_0Y_0, Y_0\rangle \geqq \underline{\lambda}(A_0)\langle Y_0, Y_0\rangle,$$

$$[u_{h_d}]^2_d = \langle Z^d, Z^d\rangle,$$

and we conclude that

$$\langle \bar{A}_d\bar{U}_d, \bar{U}_d\rangle \geqq \frac{c_1^2}{6(d+1)^2}\left\{\frac{1}{4}\underline{\lambda}(A_0)\langle Y_0, Y_0\rangle + \langle Z^d, Z^d\rangle\right\},$$

$$\underline{\lambda}(\bar{A}_d) \geqq \frac{c_1^2}{24(d+1)^2}\min\{\underline{\lambda}(A_0), 4\}. \tag{5.12}$$

Similarly,

$$\begin{aligned}\langle \bar{A}_d\bar{U}_d, \bar{U}_d\rangle &= \int_\Omega |\nabla_{h_d}\tilde{u}_{h_d}|^2\,dx\\ &\leqq \text{(by 4.1))}\\ &\leqq 4\int_\Omega |\nabla u_{h_d}|^2\,dx\\ &\leqq \text{(with (2.12))}\\ &\leqq 4c_2^2\{|\nabla r_0u_{h_d}|^2_{L^2(\Omega)} + [u_{h_d}]^2_d\}\\ &\leqq 4c_2^2\{2|\nabla_{h_0}(r_0\widetilde{u_{h_d}})|^2_{L^2(\Omega)} + [u_{h_d}]^2_d\}\\ &\leqq 4c_2^2\{2\bar{\lambda}(A_0)\langle Y_0, Y_0\rangle + \langle Z^d, Z^d\rangle\}.\end{aligned} \tag{5.13}$$

Thus

$$\bar{\lambda}(\bar{A}_d) \leqq 8c_2^2\max\{\bar{\lambda}(A_0), \tfrac{1}{2}\} \tag{5.14}$$

and

$$\kappa(\bar{A}_d) \leqq \frac{192c_2^2}{c_1^2}(d+1)^2 \frac{\max\{\bar{\lambda}(A_0), \frac{1}{2}\}}{\min\{\underline{\lambda}(A_0), 4\}}. \tag{5.15}$$

We expect $\underline{\lambda}(A_0)$ to be small (less than 4) and $\bar{\lambda}(A_0)$ to be large (larger than $\frac{1}{2}$), although the ratio

$$\kappa(A_0) = \bar{\lambda}(A_0)/\underline{\lambda}(A_0)$$

should not be too large. Hence

$$\kappa(\bar{A}_d) \lesssim \text{const}\,(d+1)^2\kappa(A_0), \tag{5.16}$$

and when d increases, *the condition number of $\bar{A}_d$ varies like d^2, in contrast with the condition number of A_d that is exponential with respect to d.* Numerical evidences are also enclosed with $\Omega = (0, 1)^2$ (see Fig. 6.1). It shows that there is a sharp difference between the condition numbers of $\bar{A}_d$ and A_d. With $d = 5$, we have 64×64 modes on the finest grid; the difference between 20.84 (condition number of $\bar{A}_d$) and 1659 (condition number of A_d) is already large.

The drawback with IUs is that $\bar{A}_d$ is a rather full matrix and direct or relaxation methods for solving (5.5) are not feasible. However, as indicated in the introduction, system (5.5) can be solved by conjugate gradient methods, in which case we only need the product of $\bar{A}_d$ with certain vectors, $\bar{A}_d\bar{U} = {}^tS_dA_dS_d\bar{U}$, and this is easy since S_d is an extremely simple matrix. We recall [AB] that k steps of the conjugate gradient method reduce the energy norm of the error $\langle \bar{A}_d\bar{U}, \bar{U}\rangle = \langle A_dU, U\rangle$, at least by the factor

$$\frac{2}{\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k},$$

and one needs at most

$$k \geqq \frac{1}{2}\sqrt{\kappa}\left|\log\left(\frac{\varepsilon}{2}\right)\right|$$

steps to reduce the norm of the error by the factor ε. For example, with 64×64 modes on $\Omega = (0, 1)^2$, we have $\kappa(\bar{A}_d) \approx 20.84$ (see Fig. 6.1). Therefore, we need only at most three steps of the conjugate gradient method to reduce the energy norm of the error by 10^{-1}. Numerical evidence of the efficiency of the IU method appears in [CT1].

5.2. More general operators. More general differential operators and more general domains Ω can be handled in the same way. For instance, consider the Dirichlet problem

$$\begin{aligned} &-\sum_{i,j=1}^{2} \frac{\partial}{\partial x_j}\left(a_{ij}(x)\frac{\partial u}{\partial x_i}(x)\right) = f(x) \quad \text{in } \Omega, \\ &u = 0 \text{ on } \partial\Omega, \end{aligned} \tag{5.17}$$

where the functions a_{11}, $a_{12} = a_{21}$, and a_{22} are given in $L^\infty(\Omega)$ and satisfy

$$\begin{aligned} &\underline{\alpha}\sum_{i=1}^{2}\xi_i^2 \leqq \sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \leqq \bar{\alpha}\sum_{i=1}^{2}\xi_i^2 \\ &\forall\xi = (\xi_1, \xi_2) \in \mathbb{R}^2 \quad \text{with } 0 < \underline{\alpha} \leqq \bar{\alpha}. \end{aligned} \tag{5.18}$$

The quadratic form associated with $\bar{A}_d$ is now

$$\langle \bar{A}_d \bar{U}_d, \bar{U}_d \rangle = \langle A_d U_d, U_d \rangle$$
$$= \sum_{i,j=1}^{2} \int_\Omega a_{ij}(x) \nabla_{i,h_d} \tilde{u}_{h_d}(x) \nabla_{j,h_d} \tilde{u}_{h_d}(x) dx.$$

Due to (5.18),

$$\langle \bar{A}_d \bar{U}_d, \bar{U}_d \rangle \geqq \underline{\alpha} \int_\Omega |\nabla_{h_d} \tilde{u}_{h_d}|^2 \, dx,$$

$$\langle \bar{A}_d \bar{U}_d, \bar{U}_d \rangle \leqq \bar{\alpha} \int_\Omega |\nabla_{h_d} \tilde{u}_{h_d}|^2 \, dx.$$

Hence, now, instead of (5.12), (5.14), and (5.16) we have

$$\underline{\lambda}(\bar{A}_d) \geqq \frac{c_1^2 \underline{\alpha}}{24(d+1)^2} \min \{ \underline{\lambda}(A_0), 4 \},$$

$$\bar{\lambda}(\bar{A}_d) \leqq 8 c_2^2 \bar{\alpha} \max \left\{ \bar{\lambda}(A_0), \frac{1}{2} \right\},$$

$$\kappa(\bar{A}_d) \lesssim \text{const}\,(d+1)^2 \frac{\bar{\alpha}}{\underline{\alpha}} \kappa(A_0).$$

The conclusions are the same.

5.3. More general domains. A more general domain $\Omega \subset \mathbb{R}^2$ can be handled in the same way. Starting as in § 2 from meshes $h_0, \dots, h_d$, and grids $\mathscr{R}_0 = \mathscr{R}_{h_0}, \dots, \mathscr{R}_d = \mathscr{R}_{h_d}$, we define

$$\mathscr{T}_d = \{ K_{j,d} \subset \Omega, \, K_{j,d} + h_i e_i \subset \Omega, \, i = 1, 2 \}.$$

Then, as before, $\mathscr{U}_d = \mathscr{R}_d \cap \bar{\Omega}$. Now the functions of V_d are only defined on

$$\Omega_d = \Omega_{h_d} = \bigcup_{K \in \mathscr{T}_d} K \, (\subset \Omega),$$

and are continuous and Q_1 on the K_s' of $\mathscr{T}_d$.

Theorem 2.1 and the results of §§ 3 and 4 are still valid, and if we solve the Dirichlet problem in Ω (in the form (5.1) or (5.17)) with incremental unknowns, we obtain the same bound on $\kappa(\bar{A}_d)$.

5.4. Neumann problem. Again let $\Omega = (0, a_1) \times (0, a_2)$. We consider the Neumann problem

$$-\Delta u + u = f \quad \text{in } \Omega, \qquad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega, \tag{5.19}$$

where ν is the outward unit normal on $\partial\Omega$.

If we use the five-point approximation of the Laplace operator and the natural approximation of the boundary condition, we obtain a system like (5.2) for which the associated bilinear form reads [Ce], [T1]

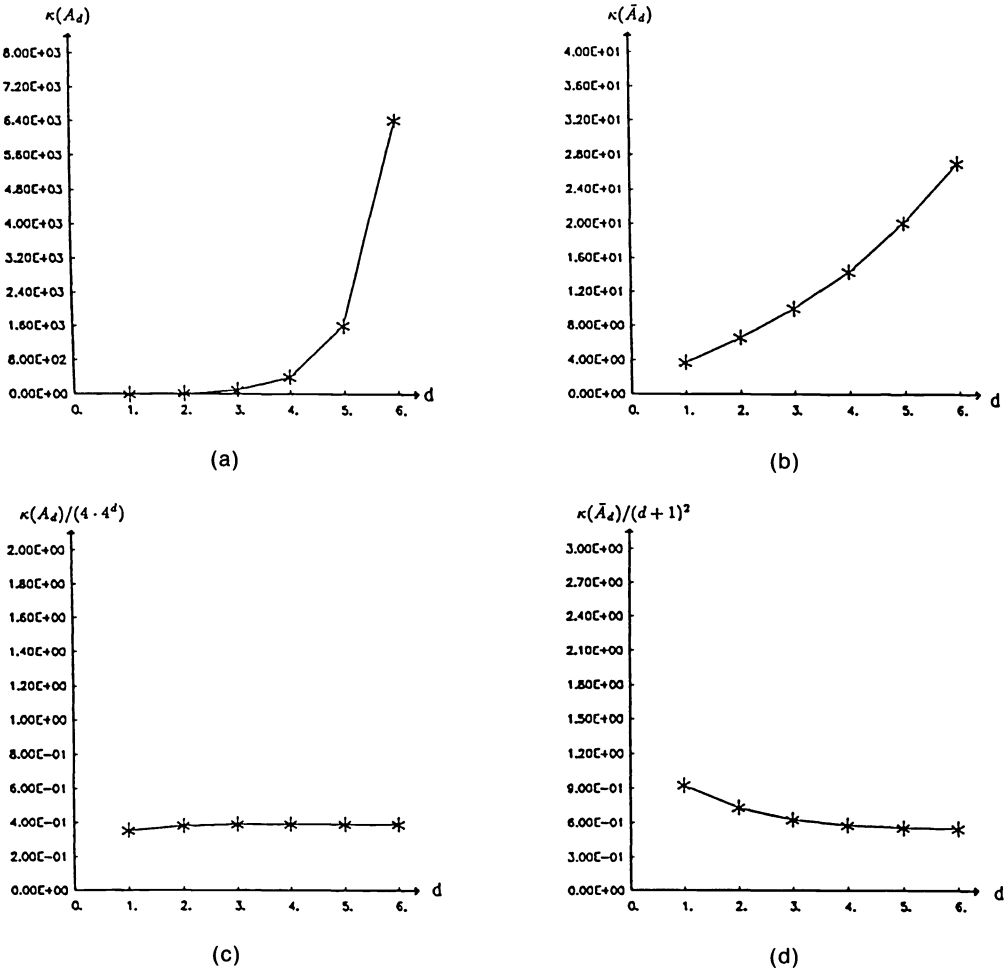
$$\langle A_d U_d, U_d^* \rangle = \int_\Omega \nabla^e_{h_d} \tilde{u}_{h_d} \nabla^e_{h_d} \tilde{u}^*_{h_d} \, dx + \int_{\Omega^*_{h_d}} \tilde{u}_{h_d} \tilde{u}^*_{h_d} \, dx, \tag{5.20}$$

where $\Omega^*_{h_d}$ is an extension of Ω; $\Omega^*_{h_d} = (0, a_1 + h_{1,d}) \times (0, a_2 + h_{2,d})$. If we implement the IUs, we obtain a system like (5.6). With the same reasoning as before, we obtain a bound similar to (5.16) for the condition number of $\bar{A}_d$: it grows quadratically with respect to d. For the proof we use (2.13) instead of (2.12), (3.9), and Lemmas 4.2 and 4.3.

6. Numerical results. We present here the numerical tests done with three examples. The condition numbers $\kappa(A_d)$ and $\kappa(\bar{A}_d)$ are compared in Figs. 6.1–6.3. They demonstrate that the condition number $\kappa(\bar{A}_d)$ is much smaller than $\kappa(A_d)$ for these operators and $\kappa(\bar{A}_d)$ grows quadratically with respect to d while $\kappa(A_d)$ grows exponentially with respect to d.

Considering the Dirichlet problem on $(0, 1)^2$ with meshes $h = 1/2^{d+1}$, $d \in \mathbb{N}$, we introduce incremental unknowns successively in d levels, as described in § 1, and leave the coarsest grid with only one unknown in the center of the square.

The first example is for the Laplace operator (see Fig. 6.1); $\kappa(A_d)$, $\kappa(\bar{A}_d)$, $\kappa(A_d)/(4 \cdot 4^d)$, and $\kappa(\bar{A}_d)/(d + 1)^2$ against d are plotted, respectively, on Figs. 6.1(a)–(d).



(a)

(b)

(c)

(d)

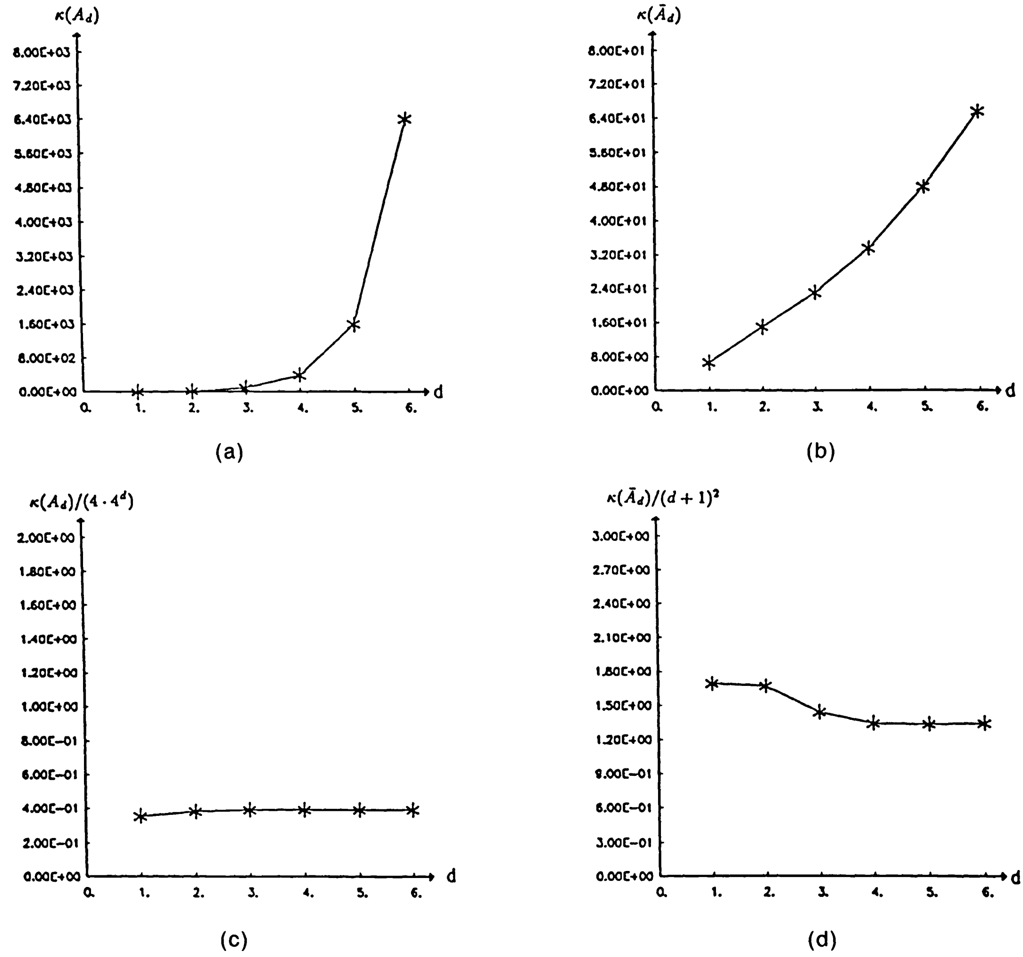
FIG. 6.1. *Comparison between the condition numbers $\kappa(A_d)$ and $\kappa(\bar{A}_d)$. A_d is the matrix in the five-point finite difference equation associated with the Laplace operator.*

Figures 6.1(a) and (b) show that the difference between $\kappa(A_d)$ and $\kappa(\bar{A}_d)$ is very large. We can see quantitatively that when $d = 6$, $\kappa(A_d) \simeq 6400$ while $\kappa(\bar{A}_d) \approx 28$. Figure 6.1(c) shows that $\kappa(A_d)/(4\cdot 4^d)$ is almost a horizontal line and is bounded above by 1. This confirms that $\kappa(A_d)$ increases with the speed of 4^d, that is, increases exponentially with respect to d. In this very special case, it is well known that the eigenvalues of $\kappa(A_d)$ are $4/h^2[\sin^2 \frac{\kappa\pi h}{2} + \sin^2 \frac{m\pi h}{2}]$, $k, m = 1, 2, \ldots, N-1$, where $N = 1/h$. Therefore, the condition number $\kappa(A_d)$ is $1 + (1/\sin^2(\pi h/2))$, which satisfies

$$\frac{4}{\pi^2}(4\cdot 4^d) \leqq \kappa(A_d) \leqq 4\cdot 4^d.$$

Compared with Fig. 6.1(c), Fig. 6.1(d) shows that $\kappa(\bar{A}_d)/(d+1)^2$ is bounded above by 2. This coincides with the results in Theorem 2.1.

The second and third examples are for the anisotropic operators. We take $a_{11} = 10$, $a_{12} = a_{21} = 0$, $a_{22} = 1$ in (5.17) for the second example and $a_{11} = 10^{x+y-1}$, $a_{12} = a_{21} = 0$, $a_{22} = 1$ in (5.17) for the third example. The conclusions are the same as for the first example and the numerical results are shown in Figs. 6.2 and 6.3, respectively. It is also



(a)

(b)

(c)

(d)

FIG. 6.2. *Comparison between the condition numbers $\kappa(A_d)$ and $\kappa(\bar{A}_d)$. A_d is the matrix in the five-point finite difference equation associated with the operator $-10\partial^2 u/\partial^2 x - \partial^2 u/\partial^2 y = f$.*

interesting to note that $\kappa(A_d)$ in the second example has exactly the same value as in the first example, but the values of $\kappa(\bar{A}_d)$ are different.

It is shown numerically in [CT1] that for Laplace operator, the efficiency of the IU method is comparable with the standard V-cycle multigrid method. For the anisotropic operators (e.g., Examples 2 and 3), the efficiency of the standard multigrid method with uniform meshes is reduced. It is usually necessary to take different relaxation processes and/or different coarsing in each case to ensure the efficiency of the method [Mc]. But for our incremental unknown method, the only efficiency factor is the condition number $\kappa(\bar{A}_d)$. Since $\kappa(\bar{A}_d) \leqq \text{const}\,(d+1)^2$ for all of the three operators above and the constants are quite small (see Figs. 6.1(d), 6.2(d), and 6.3(d)), we can expect good efficiency results by using the IU method.

Note added in proofs. It was incorrectly stated in a review of [CT1] (*Math. Reviews*, 92e: 65133, 1992, p. 2794), that the incremental unknowns method is identical to the multilevel splitting method in the context of hierarchical bases in finite elements. Some further comments seem necessary to avoid misinterpretation of our results. The rela-

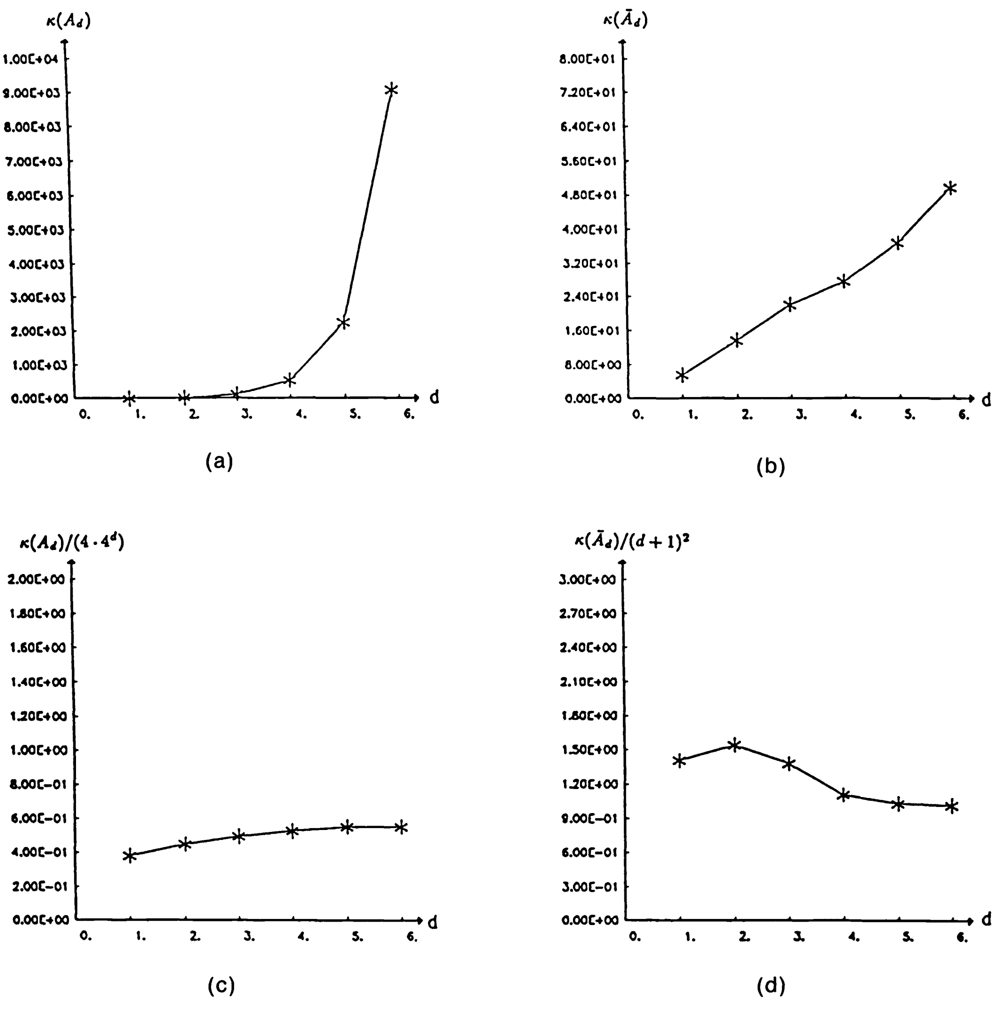


FIG. 6.3. *Comparison between the condition numbers $\kappa(A_d)$ and $\kappa(\bar{A}_d)$. A_d is the matrix in the five-point finite difference equation associated with the operator $-\partial(a(x,y)\partial u)/\partial^2 x - \partial^2 u/\partial^2 y = f$, $a(x,y) = 10^{(x+y-1)}$.*

tionship of IUs to other methods in linear algebra is explained in the Introduction of this article. Concerning the more specific case of hierarchical bases finite elements, it is not possible to identify the IU method with this method. The analogy exists: the IUs studied in this article (second-order IUs) are the same as the components of the function in the $Q1$-hierarchical basis; this was mentioned before and we indeed rely on the corresponding results available in the literature. However, the differences are numerous. The matrices of the linear system are not the same in general for finite differences and finite elements. There is also the difference in implementation with, in this regard, the advantages and drawbacks of both methods. IUs are a much more flexible and general concept and in [CT2] we actually introduce and study IUs that are not related to finite elements and are, in fact, in some cases related to wavelets.

Finally, we recall that the ultimate goal with IUs is the study of nonlinear problems when turbulent phenomena occur. Theoretically and computationally new interesting phenomena occur in this direction, which seems to be a promising new chapter in the numerical analysis of dissipative evolution equations.

Indeed the new technologies and the increased power of the new computers offer new challenging problems to the numerical analysts, namely, the approximation of nonlinear evolution equations on large intervals of time when complex physical phenomena appear. New numerical methods adapted to such problems need to be developed; in particular, multilevel methods are needed in order to treat appropriately the different scales appearing in a complex problem and to resolve at reasonable cost the smaller scales. The approximation of time evolution equations in the presence of chaotic phenomena is the source of many new challenging problems in numerical analysis, which have been barely touched on (see, e.g., [ChT], [CT3], [ES], [FJKT], [Sh], [T3]).

REFERENCES

[AB] O. Axelsson and V. A. Baker, *Finite Element Solution of Boundary Value Problems: Theory and Computation*, Academic Press, New York, 1984.

[AS] J. Atanga and D. Silvester, *Preconditioning techniques for the numerical solution of the Stokes problem*, in Proc. Second Internat. Conference on the Application of Supercomputers in Engineering (ASE91), Boston, August 1991, to appear.

[B] D. Braess, *The contraction number of a multigrid method for solving the Poisson equation*, Numer. Math., 37 (1981), pp. 387-404.

[BVY] G. Birkhoff, R. S. Varga, and D. Young, *Alternating implicit methods*, in Advances in Computers, Vol. 3, F. Alt and M. Rubinoff, eds., Academic Press, New York, 1962.

[C] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1988.

[Ce] J. Céa, *Approximation variationnelle des problèmes aux limites*, Ann. Inst. Fourier, 14 (1964), pp. 345-444.

[ChT] H. Choi and R. Temam, *Application of incremental unknowns to the Burgers equation*, in Advances in Computational Fluid Dynamics, M. Hafez, ed., to appear.

[CT1] M. Chen and R. Temam, *Incremental unknowns for solving partial differential equations*, Numer. Math., 59 (1991), pp. 255-271.

[CT2] ———, *Nonlinear Galerkin method in finite difference case and wavelet-like incremental unknowns*, Numer. Math., 533 (1992), to appear.

[CT3] ———, *Nonlinear Galerkin method with multilevel incremental unknowns*, in Contributions in Numerical Mathematics, Volume in memory of Prof. Collatz, R. P. Agarwal, ed., World Scientific Series in Applicable Analysis, Singapore, 1993, to appear.

[DW] M. Dryja and O. B. Widlund, *Multilevel Additive Methods for Elliptic Finite Element Problems*, preprint, 1991.

[ES] C. M. Elliott and A. M. Stuart, *The global dynamics of discrete semilinear parabolic equations*, IMA Preprint Series No. 973, Univ. of Minnesota, Minneapolis, MN, 1992.

[FJKT] C. Foias, M. S. Jolly, J. G. Kevrekidis, and E. S. Titi, *Dissipativity of numerical schemes*, Nonlinearity, 4 (1991), p. 591-613.

[FMT] C. Foias, O. Manley, and R. Temam, *On the interaction of small and large eddies in two-dimensional turbulent flows*, Math. Modelling and Numer. Anal., 22 (1988), pp. 93–114.

[FMT2] C. Foias, O. Manley, R. Temam, and Y. Treve, *Asymptotic analysis of the Navier–Stokes equations*, Phys. D, 9D (1983), pp. 157–188.

[GM] G. H. Golub and G. A. Meurant, *Résolution Numérique des Grands Systèmes Linéaires*, Collection DER-EDF, Vol. 49, Eyrolles, Paris, 1983.

[GTr] D. Gilberg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Heidelberg, New York, 1983.

[GV] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.

[Mc] S. F. McCormick, ed., *Multigrid Methods*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987.

[MT] M. Marion and R. Temam *Nonlinear Galerkin methods; the finite elements case*, Numer. Math., 57 (1990), pp. 205–226.

[S] G. Strang, *Linear Algebra and its Applications*, Academic Press, New York, 1976.

[Sh] J. Shen, *Convergence of approximate attractors for a fully discrete system for reaction-diffuse systems*, Numer. Funct. Anal. Optim., 10 (1989), p. 1213–1234.

[T1] R. Temam, *Numerical Analysis*, Reidel, Dordrecht, 1973.

[T2] ———, *Inertial manifolds and multigrid methods*, SIAM J. Math. Anal., 21 (1990), pp. 154–178.

[T3] ———, *Stability analysis of the nonlinear Galerkin method*, Math. Comp., 57 (1991), p. 477–505.

[Th] V. Thomée, *Galerkin finite element methods for parabolic problems*, Lecture Notes in Math. 1054, Springer-Verlag, Berlin, Heidelberg, New York, 1982.

[W] W. L. Wendland, *Elliptic Systems in the Plane*, Pitman, London, San Francisco, Melbourne, 1979.

[X1] J. Xu, *Theory of multilevel methods*, Ph. D. thesis, Cornell Univ., Ithaca, NY, 1989; Tech. Rep. AM-48, Mathematics Dept., Pennsylvania State Univ., State College, PA, 1989.

[X2] ———, *Iterative methods by space decomposition and subspace correction: A unifying approach*, Tech. Rep. AM-67, Pennsylvania State Univ., State College, PA, 1990.

[X3] ———, *A new class of iterative methods for nonselfadjoint or indefinite problems*, Tech. Rep., Mathematics Dept., Pennsylvania State Univ., State College, PA, 1991.

[Y1] H. Yserentant, *On the multi-level splitting of finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.

[Y2] ———, *Two preconditioners based on the multi-level splitting of finite elements*, Numer. Math., 58 (1990), pp. 163–184.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 456–467, April 1993

010

ON ACCURATE COMPUTATIONS OF THE PERRON ROOT*

L. ELSNER†, I. KOLTRACHT‡§, M. NEUMANN‡¶, AND D. XIAO‡#

Abstract. This paper establishes a new componentwise perturbation result for the Perron root of a nonnegative and irreducible matrix. The error bound is independent of the angle between left and right Perron eigenvectors. It is shown that a known inverse iteration algorithm with new stopping criteria will have a small componentwise backward error, which is consistent with the perturbation result. Numerical experiments demonstrate that the accuracy of the Perron root computed by the proposed algorithm is, indeed, independent of the angle.

Key words. nonnegative matrices, Perron root, sparse systems, backward error, componentwise perturbations, stable algorithms

AMS(MOS) subject classifications. 65F70, 15A06, 15A18

1. Introduction. In this paper we consider the problem of how to accurately compute the Perron root of a nonnegative and irreducible matrix A. It is well known (see Berman and Plemmons [3]) that the Perron root, which equals the spectral radius of A, is a simple eigenvalue to which there corresponds a positive eigenvector.

For general matrices, the sensitivity of a simple eigenvalue to perturbations depends on the angle between normalized left and right eigenvectors corresponding to the eigenvalue. The following result appears in Wilkinson's book [10]: *If λ is a simple eigenvalue of a square matrix A and y and x are corresponding normalized left and right eigenvectors, then for sufficiently small δ the matrix $A + \delta E$, with $\|E\|_2 = 1$, has a simple eigenvalue $\hat{\lambda}$ that satisfies the inequality*

$$|\lambda - \hat{\lambda}| \leqq \frac{\delta}{|y^T x|} + o(\delta).$$

Using a formula for the componentwise condition number for general continuous maps given in Gohberg and Koltracht [6], we shall obtain in § 2 the componentwise condition number of a simple eigenvalue of an $n \times n$ matrix A and compare it with the usual normwise condition number. In the special case when A is an $n \times n$ nonnegative and irreducible matrix and the eigenvalue in question is the Perron root, then for relatively small componentwise perturbations in A, i.e.,

$$(1.1) \qquad |E_{i,j}| \leqq \varepsilon A_{i,j}, \qquad i, j = 1, \ldots, n,$$

* Received by the editors August 7, 1991; accepted for publication (in revised form) November 12, 1991.

† Department of Mathematics, University of Bielefeld, P.O. Box 8640, 4800-Bielefeld 1, Germany. The research of this author was supported in part by Sonderforschungsbereich 343, "Discrete Structures in Mathematics."

‡ Department of Mathematics, University of Connecticut, Storrs, Connecticut 06269-3009.

§ The research of this author was supported in part by National Science Foundation grant DMS-9007030 and by Sonderforschungsbereich 343, "Discrete Structures in Mathematics," through the Department of Mathematics, University of Bielefeld.

¶ The research of this author was supported in part by U.S. Air Force Office of Scientific Research grant AFOSR-88-0047, by National Science Foundation grants DMS-8901860 and DMS-9007030, and by Sonderforschungsbereich 343, "Discrete Structures in Mathematics," through the Department of Mathematics, University of Bielefeld.

# The research of this author was supported in part by National Science Foundation grants DMS-8901860 and DMS-9007030.

we shall show that

$$\frac{|\hat{\lambda} - \lambda|}{\lambda} \leqq \varepsilon + o(\varepsilon). \tag{1.2}$$

Hence the componentwise condition number of the Perron root is 1. In fact, we show in Theorem 1 of § 2 that a better result than (1.2) is possible. It is that

$$\frac{|\hat{\lambda} - \lambda|}{\lambda} \leqq \varepsilon. \tag{1.3}$$

In view of (1.3) a natural question to ask is whether one can find an algorithm for computing the Perron root that is numerically stable in the sense of (1.1). To be precise, we are looking for an algorithm for computing the Perron root of A for which the computed root is the exact Perron root of a perturbation $A + E$ of A satisfying

$$|E_{i,j}| \leqq f(n)uA_{i,j}, \qquad i, j = 1, \ldots, n, \tag{1.4}$$

where u is the unit round-off error and $f(n)$ is a slowly growing function. Then the computed root will satisfy

$$\frac{|\hat{\lambda} - \lambda|}{\lambda} \leqq f(n)u. \tag{1.5}$$

Since any algorithm for computing λ is iterative, we must find an algorithm such that, subject to certain stopping criteria, the approximate root produced by the algorithm will be an exact root of $A + E$, where E is small and satisfies (1.4).

One such algorithm is presented in § 3. It is based on a variant of the inverse iteration due to Noda [7], which was shown by Elsner [4] to be quadratically convergent to the Perron root and which involves, at each stage of the iteration, the solution of a linear system whose coefficient matrix is an M-matrix. Using a theorem of Skeel [9], we show that if at each stage of the algorithm the M-matrix system is solved by Gaussian elimination followed by an iterative refinement of the solution and if certain stopping criteria are satisfied, then this results in an algorithm such that the approximation to the Perron root of A is an exact Perron root of $A + E$ with E satisfying (1.4). This result is given in § 3.

In § 4 we present some numerical experiments that illustrate the stability of the proposed algorithm. In particular, we give an example of an M-matrix for which the standard QR algorithm breaks down, whereas the algorithm of § 3 computes the Perron root with high accuracy, as expected. These results also show that the power-squaring algorithm (see Friedland [5])

$$\lambda = \rho(A) = \lim_{k \to \infty} \|A^{2^k}\|_{\infty}^{1/2^k}, \tag{1.6}$$

where $\rho(\cdot)$ denotes the spectral radius, which is seemingly natural from the point of view of round-off error analysis, is inferior in its convergence speed. (However, it should be remarked that on a vector or parallel computer, the power-squaring method may become competitive because its rate of convergence could be compensated for by fast matrix-matrix multiplications.)[1] Our numerical experiments also show that the computed Perron vectors are highly accurate, although we do not yet have a full perturbation analysis for explaining this evidence.

[1] Friedland [5] also notes that the power-squaring algorithm could be used to compute the subdominant eigenvalue of a nonnegative and irreducible matrix.

2. Perturbation results. Suppose that the matrix $A \in R^{n,n}$ has a simple eigenvalue $\lambda = \lambda(A)$. It is well known that the map $F_E : \varepsilon \rightarrow \lambda(A + \varepsilon E)$ is analytic in a neighborhood of 0; see, for example, Wilkinson [10, pp. 66–67]. Therefore, the map $F : \hat{A} \rightarrow \lambda(\hat{A})$ has continuous partial derivatives with respect to each entry at A and (see [10, Chap. 2])

$$\frac{\partial F}{\partial_{i,j}}(A) := \lim_{t \rightarrow 0} \frac{F(A + tE_{i,j}) - F(A)}{t} = \frac{y_i x_j}{y^T x}, \qquad i, j = 1, \ldots, n.$$

Hence, as a map from $R^{n^2} \rightarrow R$, F is differentiable at A and

$$F'(A) = \left[\frac{\partial F}{\partial_{1,1}}, \ldots, \frac{\partial F}{\partial_{1,n}}, \frac{\partial F}{\partial_{2,1}}, \ldots, \frac{\partial F}{\partial_{n,n}}\right].$$

According to a formula of Gohberg and Koltracht [6], the sensitivity of $F(A)$ to componentwise perturbations in A as in (1.1), for $\lambda \neq 0$, is characterized by the componentwise condition number of F at A given by

$$c(F, A) = \frac{\|F'(A)D_A\|_\infty}{\|F(A)\|_\infty} = \frac{\|F'(A)D_A\|_\infty}{|\lambda|},$$

where

$$D_A = \operatorname{diag}(a_{1,1}, \ldots, a_{1,n}, a_{2,1}, \ldots, a_{n,n}).$$

This means that

$$c(F, A) = \frac{\|(y_1 a_{1,1} x_1, \ldots, y_1 a_{1,n} x_n, y_2 a_{2,1} x_1, \ldots, y_n a_{n,n} x_n)\|_\infty}{|\lambda|},$$

where the infinity norm of the map $F'(A)D_A$ is, in fact, the 1-norm of the row vector that represents it. Thus

$$c(F, A) = \frac{\sum_{i,j=1}^n |y_i a_{i,j} x_j|}{|\lambda|\,|y^T x|} = \frac{|y|^T |A|\,|x|}{|\lambda|\,|y^T x|}, \tag{2.1}$$

where an absolute value of a matrix is the corresponding matrix of the absolute values of its entries. Thus if E is a perturbation of A which satisfies (1.1), the $\hat{\lambda} = \lambda(A + E)$ satisfies

$$\frac{|\hat{\lambda} - \lambda|}{|\lambda|} \leqq c(F, A)\varepsilon + o(\varepsilon), \tag{2.2}$$

where $o(\varepsilon)$ means that $\lim_{\varepsilon \rightarrow 0} / \varepsilon = 0$.

It is interesting to compare (2.1) with the usual condition number taken with respect to normwise perturbations in A. For example, let us consider perturbations in A with respect to the Frobenius norm, in which case

$$k(F, A) = \frac{\|F'(A)\|_2 \|A\|_F}{|\lambda|},$$

where the 2-norm of $F'(A)$ is its norm as a map from R^{n^2} to R and therefore coincides with its usual 2-norm. Since $\|x\|_2 = \|y\|_2 = 1$, it is clear that

$$\|F'(A)\|_2 = \frac{1}{|y^T x|} \sum_{i,j=1}^n (x_i y_j)^2 = \frac{1}{|y^T x|},$$

and so

$$k(F, A) = \frac{\|A\|_F}{|\lambda|\,|y^T x|}. \tag{2.3}$$

It is clear that $c(F, A) \leqq k(F, A)$. Moreover, it is possible that $k(F, A)$ is large, while $c(F, A)$ is much smaller. Indeed, this may be the case if the smallness of $y^T x$ and/or λ is offset by a small $|y^T|\,|A|\,|x|$.

Suppose now that A is a nonnegative and irreducible matrix and λ is its Perron root. Then, because x and y are positive vectors, $|y^T|\,|A|\,|x| = y^T A x = \lambda y^T x$, and we see at once that $c(F, A) = 1$. According to (2.2),

$$\frac{|\hat{\lambda} - \lambda|}{|\lambda|} \leqq \varepsilon + o(\varepsilon).$$

We can, in fact, improve on this result.

THEOREM 1. *Suppose that A is an $n \times n$ nonnegative and irreducible matrix, and suppose that E is an $n \times n$ real matrix such that*

$$|E| \leqq \varepsilon A,$$

where $\varepsilon \leqq 1$. Let λ and λ_E denote, respectively, the Perron roots of A and $A + E$. Then

$$\frac{|\lambda_E - \lambda|}{\lambda} \leqq \varepsilon. \tag{2.4}$$

Proof. The inequality (1.1) can be written as

$$0 \leqq A - \varepsilon A \leqq A + E \leqq A + \varepsilon A.$$

Since $\rho(\cdot)$ is monotone on the nonnegative matrices (see, for example, [3]), it follows that

$$\rho(A - \varepsilon A) \leqq \rho(A + E) \leqq \rho(A + \varepsilon A).$$

Since $\rho(A \pm \varepsilon A) = (1 \pm \varepsilon)\rho(A)$, we get that

$$(1 - \varepsilon)\lambda \leqq \lambda_E \leqq (1 + \varepsilon)\lambda.$$

Because $\lambda > 0$, the last inequality is equivalent to (2.4). □

3. A stable algorithm for computing the Perron root.

Let A be an irreducible nonnegative matrix, p be the Perron vector of A whose infinity norm is 1, and $\rho(A)$ be the Perron root of A. The basis for our algorithm is a certain inverse iteration due to Noda [7], which has been generalized and shown to be quadratically convergent by Elsner [4].

INVERSE ITERATION ALGORITHM. For a given $y_0 > 0$ define iteratively

$$\mu_s = \max\left(\frac{Ay_s}{y_s}\right),$$

$$x_s = (\mu_s I - A)^{-1} y_s,$$

and

$$y_{s+1} = \frac{x_s}{\|x_s\|_\infty}.$$

Then, for any s, $\rho(A) \leqq \mu_{s+1} \leqq \mu_s$ and

$$\mu_{s+1} - \rho(A) \leqq C_1(\mu_s - \rho(A))^2$$

and

$$\operatorname{osc}\left(\frac{y_{s+1}}{p}\right) \leqq C_2\left[\operatorname{osc}\left(\frac{y_s}{p}\right)\right]^2,$$

where C_1 and C_2 are some constants depending on y_0 and A only, and where, for any two n-vectors $u = (u_1, \ldots, u_n)^T$ and $v = (v_1, \ldots, v_n)^T$, $v > 0$, by definition

$$\max\left(\frac{u}{v}\right) = \max_{1 \leqq i \leqq n} \frac{u_i}{v_i}, \qquad \min\left(\frac{u}{v}\right) = \min_{1 \leqq i \leqq n} \frac{u_i}{v_i},$$

and

$$\operatorname{osc}\left(\frac{u}{v}\right) = \max\left(\frac{u}{v}\right) - \min\left(\frac{u}{v}\right).$$

(For more background material concerning properties and applications of vector *oscillation*, see Seneta [8, Chap. 3.4].) Since $\mu_s > \rho(A)$ so long as y_0 is not a scalar multiple of p, it follows that the matrix $(\mu_s I - A)$ is an M-matrix. A certain version of the Gaussian elimination algorithm due to Ahac, Buoni, and Olesky [2] can be used to compute the solution of the equation $(\mu_s I - A)x_s = y_s$. According to [2], the pivots will always be found among the diagonal elements, and this saves some execution time. It is also shown in [2] that the computed solution is exact for a perturbed system whose relative distance from the original system is $O(n)u$, where u is the machine precision. This distance, however, is measured normwise, whereas the bound of Theorem 1 requires componentwise perturbations. In order to get a small componentwise backward error, we add a one-step iterative refinement, as recommended by the following theorem of Skeel [9] concerning solution of $Ax = b$ by Gaussian elimination.

THEOREM 2 (Skeel [9]). *Let u be the machine precision, and let the arithmetic be such that the floating-point result* fl $(a*b)$ *of the operation $a*b$, where $* \in \{+, -, \times, /\}$, satisfies* fl $(a*b) = (a*b)(1 + e)$ *with $|e| \leqq u$. There is a function $f(A, b)$, typically behaving as $O(n)$, such that when the product of $\hat{k}(A) = \| |A| |A^{-1}| \|$ and $\sigma(A, x) = \max(|A| |x|)/\min(|A| |x|)$ is less than $(f(A, b)u)^{-1}$ and there is no overflow or underflow, then the following one-step refinement*:

Solve $Ax = b$ using Gaussian elimination, obtaining solution $\hat{x}$, and saving the LU *factors*;

Compute the residual $r = A\hat{x} - b$ (using single-precision u);

Solve $Ad = r$ for d using the saved LU *factors of A*;

Update $\hat{x} = \hat{x} - d$;

gives a vector $\hat{x}$ that is the exact solution of the equation

$$(A + \Delta A)\hat{x} = b + \Delta b,$$

where

$$|\Delta A_{i,j}| \leqq (n+1)u|A_{i,j}|, \qquad i, j = 1, \ldots, n,$$

and

$$|\Delta b_j| \leqq (n+1)u|b_j|, \qquad j = 1, \ldots, n.$$

We formulate the following algorithm for computing approximations to the Perron root $\rho(A)$ and the Perron vector p of a nonnegative irreducible matrix A.

ALGORITHM. Let u be the machine precision. Start with $y_0 > 0$ and $\mu_0 = \max(Ay_0/y_0)$. For $s = 0, 1, \ldots$

1. Compute the LU factorization

$$(\mu_s I - A) = L_s U_s,$$

and solve for x_s

$$L_s U_s x_s = y_s$$

by the Ahac, Buoni, and Olesky algorithm; save the LU factors;

2. Compute $r = Ax_s - y_s$;
3. Solve $Ad = r$ using the saved LU factors L_s and U_s;
4. Update $\bar{x}_s = x_s - d$;
5. Compute

$$y_{s+1} = \frac{\bar{x}_s}{\|\bar{x}_s\|_\infty} \quad \text{and} \quad \mu_{s+1} = \max\left(\frac{Ay_{s+1}}{y_{s+1}}\right);$$

6. Proceed until

$$\|\bar{x}_s\|^{-1} \leqq u^{1/2} \quad \text{and} \quad \operatorname{osc}\left(\frac{y_s}{y_{s+1}}\right) \leqq u^{1/2}.$$

The following theorems characterize the numerical behavior of the algorithm and explain the stopping criteria in step 6 above. Variables computed by the algorithm are denoted by hats. The norm below always means the infinity norm. It is assumed for simplicity that the matrix A is of the size $(n-1) \times (n-1)$.

THEOREM 3. *Suppose that the algorithm terminates when*

$$\|\hat{\bar{x}}_s\|^{-1} = \varepsilon_1 \quad \textit{and} \quad \operatorname{osc}\left(\frac{\hat{y}_s}{\hat{y}_{s+1}}\right) = \varepsilon_2$$

and that

$$4\eta(1+\eta) < \hat{\mu}_s \frac{1-nu}{1+nu}, \tag{3.1}$$

where $\eta = \max(\varepsilon_1, \varepsilon_2)$. *Suppose further that*

$$\|(\hat{\mu}_s I - A)^{-1} |\hat{\mu}_s I - A|\| \frac{\max(|\hat{\mu}_s - A|\hat{\bar{x}}_s)}{\min(|\hat{\mu}_s - A|\hat{\bar{x}}_s)} < (fu)^{-1}, \tag{3.2}$$

where $f = f(\hat{\mu}_s I - A, \hat{y}_s)$ *is as in Theorem* 2. *Then* $\hat{\mu}_s - \varepsilon_1$ *and* $\hat{y}_{s+1}$ *are the exact Perron root and Perron vector, respectively, of the matrix* $\bar{A}$, *where*

$$\bar{A}_{i,j} = A_{i,j}(1 + e_{i,j})(1 + \delta_i), \quad i, j = 1, \ldots, n-1, \quad i \neq j,$$

and

$$\bar{A}_{i,i} = A_{i,i}(1 + e_{i,i}), \qquad i = 1, \ldots, n-1,$$

with

$$|e_{i,j}| \leqq nu, \qquad |e_{i,i}| \leqq 2nu,$$

and

$$|\delta_i| \leqq \frac{4nu}{1-nu} + \frac{4\varepsilon_1(nu + \varepsilon_2 + nu\varepsilon_2)}{\hat{\mu}_s(1-nu)}.$$

Proof. It follows from the definition of $\hat{\hat{x}}_s$ in the algorithm and from Theorem 2 applied to the equation $(\hat{\mu}_s I - A)\hat{x}_s = \hat{y}_s$ that

$$(\hat{\mu}_s I - A - E)\hat{\hat{x}}_s = \hat{y}_s + \Delta y,$$

where

$$|E_{i,j}| \leqq nu|(\hat{\mu}_s I - A)_{i,j}| \tag{3.3}$$

and

$$|\Delta y_i| \leqq nu|(\hat{y}_s)_i|.$$

Therefore, dividing by the norm of x_s and using the definition of ε_1 we can write

$$(\hat{\mu}_s I - A - E)\hat{y}_{s+1} = \varepsilon_1(I + D_\varepsilon)\hat{y}_s, \tag{3.4}$$

where D_ε is a diagonal matrix such that $|f_i| \leqq nu$ with $D_\varepsilon(i, i) = f_i$. Note that the infinity norm of $\hat{y}_{s+1}$ and $\hat{y}_s$ equals 1, and that they are both positive vectors. Thus

$$\min\left(\frac{\hat{y}_s}{\hat{y}_{s+1}}\right) \leqq 1 \leqq \max\left(\frac{\hat{y}_s}{\hat{y}_{s+1}}\right).$$

From the equality

$$\max\left(\frac{\hat{y}_s}{\hat{y}_{s+1}}\right) = \operatorname{osc}\left(\frac{\hat{y}_s}{\hat{y}_{s+1}}\right) + \min\left(\frac{\hat{y}_s}{\hat{y}_{s+1}}\right)$$

it follows that

$$\max\left(\frac{\hat{y}_s}{\hat{y}_{s+1}}\right) \leqq 1 + \varepsilon_2 \quad \text{and} \quad \min\left(\frac{\hat{y}_s}{\hat{y}_{s+1}}\right) \geqq 1 - \varepsilon_2.$$

Therefore,

$$(1 - \varepsilon_2)\hat{y}_{s+1} \leqq \hat{y}_s \leqq (1 + \varepsilon_2)\hat{y}_{s+1},$$

and hence we can write $\hat{y}_s = (I + D_{\varepsilon_2})\hat{y}_{s+1}$, where D_{ε_2} is a diagonal matrix such that $|h_i| \leqq \varepsilon_2$ with $D_{\varepsilon_2}(i, i) = h_i$. Substituting into (3.4), we get

$$(\hat{\mu}_s I - A - E)\hat{y}_{s+1} = \varepsilon_1(1 + D_\varepsilon)(1 + D_{\varepsilon_2})\hat{y}_{s+1}. \tag{3.5}$$

This, in turn, can be rewritten as

$$\sum_{j=1, j\neq i}^{n} (A_{i,j} + E_{i,j})\hat{y}_{s+1}(j) = ((\hat{\mu}_s - A_{i,i}) - E_{i,i} - \varepsilon_1(1 + f_i)(1 + h_i))\hat{y}_{s+1}(i). \tag{3.6}$$

Let us now define $e_{i,j} = E_{i,j}/A_{i,j}$ for $i \neq j$ (where $0/0 = 0$). Let $e_{i,i} = 0$ if $A_{i,i} < \hat{\mu}_s/2$, and let $e_{i,i} = E_{i,i}/A_{i,i}$ if $A_{i,i} \geqq \hat{\mu}_s/2$. For the case $A_{i,i} < \hat{\mu}_s/2$, let δ_i be such that

$$\sum_{j=1, j\neq i}^{n} (A_{i,j} + E_{i,j})\hat{y}_{s+1}(j)\delta_i = (-E_{i,i} - \varepsilon_1(f_i + h_i + f_i h_i))\hat{y}_{s+1}(i). \tag{3.7}$$

In this case it follows from (3.6) and (3.3) that

$$\begin{aligned}
|\delta_i| &= \left|\frac{-E_{i,i} - \varepsilon_1(f_i + h_i + f_i h_i)}{(\hat{\mu}_s - A_{i,i}) - E_{i,i} - \varepsilon_1(1 + f_i)(1 + h_i)}\right| \\
&\leqq \frac{|E_{i,i}|}{\hat{\mu}_s(1 - nu)/2 - \hat{\mu}_s(1 - nu)/4} + \frac{\varepsilon_1(nu + \varepsilon_2 + nu\varepsilon_2)}{\hat{\mu}_s(1 - nu)/2 - \hat{\mu}_s(1 - nu)/4} \\
&\leqq \frac{4nu}{(1 - nu)} + \frac{\varepsilon_1(nu + \varepsilon_2 + nu\varepsilon_2)}{\hat{\mu}_s(1 - nu)/4}.
\end{aligned}$$

If $A_{i,i} \geqq \hat{\mu}_s/2$, then let δ_i be such that

$$\sum_{j=1}^{n} (A_{i,j} + E_{i,j})\hat{y}_{s+1}(j)\delta_i = -\varepsilon_1(f_i + h_i + f_i h_i)\hat{y}_{s+1}(i). \tag{3.8}$$

Again, it follows from (3.6) and (3.3) that

$$|\delta_i| = \left|\frac{-\varepsilon_1(f_i + h_i + f_i h_i)}{\hat{\mu}_s - \varepsilon_1(1 + f_i)(1 + h_i)}\right| \leqq \frac{\varepsilon_1(nu + \varepsilon_2 + nu\varepsilon_2)}{\hat{\mu}_s - \hat{\mu}_s(1 - nu)/4}.$$

Using the definition of δ_i and $e_{i,j}$ and the equality (3.5), we finally obtain

$$\sum_{j=1, j \neq i}^{n} (A_{i,j} + E_{i,j})(1 + \delta_i)\hat{y}_{s+1}(j) + A_{i,i}(1 + e_{i,i})\hat{y}_{s+1}(i) = (\hat{\mu}_s - \varepsilon_1)\hat{y}_{s+1}(i). \tag{3.9}$$

The theorem is proved. □

It follows from Theorem 1 that if $\eta = u^{1/2}$ and if the conditions (3.1) and (3.2) are satisfied, then $\hat{\mu}_s - \varepsilon_1$ is the exact Perron root of $\bar{A}$ such that $|A - \bar{A}| \leqq (u^{1/2} + O(u))A$. Indeed, it is easy to see from (3.1) that $|\delta_i| \leqq u^{1/2} + O(u)$. Thus, by Theorem 1, $|\hat{\mu}_s - \varepsilon_1 - \rho(A)| \leqq (u^{1/2} + O(u))\rho(A)$, and hence $|\hat{\mu}_s - \rho(A)| \leqq (1 + \rho(A))u^{1/2} + O(u)$. Since the sequence $\{\hat{\mu}_s\}$ converges quadratically to $\rho(A)$, it follows that

$$\hat{\mu}_{s+1} - \rho(A) \leqq C_1(1 + \rho(A))^2 u + o(u),$$

where C_1 is a constant defined in the inverse iteration algorithm.

We comment that condition (3.1) is not restrictive since A can be replaced, e.g., by $I + A$. Let us argue that the condition (3.2) is not restrictive either. This condition is necessary to assure that the product of $\hat{k}(\mu_s I - A)$ and $\sigma(\mu_s I - A, x_s)$ is not too large relative to the machine precision for Skeel's theorem to apply. However, as was pointed out by Wilkinson [10, § 9.47], it is typical for the inverse iteration that the solution of $(\mu_s I - A)x_s = y_s$ is computed with high relative accuracy, while the coefficient matrix may be very ill conditioned. Further justification of this point will be given in Theorem 4 below. The second factor, $\sigma(\mu_s I - A, x_s)$, is necessary in Skeel's theorem to account for possible zeros of the solution vector and the right-hand side, as is discussed in detail in Arioli, Demmel, and Duff [1]. Since in our case both x_s and y_s are positive vectors, one can expect Skeel's result to hold even if the factor $\sigma(\mu_s I - A, x_s)$ is large. These arguments are supported by numerical evidence in § 4. The experiments also suggest that if only the Perron root, but not the Perron vector, needs to be calculated, then the stopping criterion of the algorithm should be changed to $\|x_s\|^{-1} < u^{1/2}$, since the second criterion, which assures accuracy of the Perron vector, may slow down the algorithm to some degree. In this case, as can be seen from the proof of the theorem, condition (3.1) can be replaced by the condition

$$4\varepsilon_1(1 + \varepsilon_2) < \hat{\mu}_s \frac{1 - nu}{1 + nu}.$$

It is now clear that ε_2 does not need to be very small for this condition to be satisfied.

In the next theorem we show when the condition (3.2) will always be satisfied.

THEOREM 4. *Let $\hat{y}_{s+1}$ and $\hat{\mu}_s$ denote the quantities computed in the algorithm for some value of s. Let p be the Perron vector of A, and define*

$$S_p = \frac{\max(p)}{\min(p)}, \qquad S_y = \frac{\max(\hat{y}_{s+1})}{\min(\hat{y}_{s+1})},$$

and

$$\delta_m = \min_{1 \leqq i \leqq n} \sum_{i \neq j} A_{i,j}.$$

Then

$$\|(\hat{\mu}_s I - A)^{-1}|(\hat{\mu}_s I - A)|\|(\hat{\mu}_s - \rho(A)) \leqq 2S_p\hat{\mu}_s, \tag{3.10}$$

$$\frac{\max(|\hat{\mu}_s I - A|\,\hat{y}_{s+1})}{\min(|\hat{\mu}_s I - A|\,\hat{y}_{s+1})}\,\delta_m \leqq 2S_y\hat{\mu}_s. \tag{3.11}$$

Proof. Observe that

$$|(\hat{\mu}_s I - A)| = A - 2D_A + \hat{\mu}_s I,$$

where $D_A = \operatorname{diag}(A_{1,1}, \dots, A_{n,n})$. Let $y = |\hat{\mu}_s I - A|\hat{y}_{s+1}$. Then clearly

$$y \leqq \hat{\mu}_s \hat{y}_{s+1} + A\hat{y}_{s+1}$$

and

$$y \geqq (\hat{\mu}_s I - D_A)\hat{y}_{s+1} + \delta_m \min(\hat{y}_{s+1}).$$

It follows from the inequality $\hat{\mu}_s \geqq \hat{\mu}_{s+1}$ and the definition of $\hat{\mu}_{s+1}$ that

$$\frac{\max(y)}{\min(y)} \leqq \frac{\hat{\mu}_s + \hat{\mu}_{s+1}}{\min_i(\hat{\mu}_s - A_{i,i}) + \delta_m} S_y \leqq \frac{2\hat{\mu}_s}{(\delta_m)} S_y.$$

Thus (3.11) is proved. To show (3.10), observe that for any $\mu > \rho(A)$,

$$(\mu I - A)^{-1}p = (\mu - \rho(A))^{-1}p.$$

Let $D_p = \operatorname{diag}(p(1), \dots, p(n))$ and $e = (1, \dots, 1)^T$. Then

$$D_p^{-1}(\mu I - A)^{-1}D_p e = (\mu - \rho(A))^{-1}e,$$

which implies that

$$\|D_p^{-1}(\mu I - A)^{-1}D_p\| = (\mu - \rho(A))^{-1}.$$

Therefore,

$$\|(\mu I - A)^{-1}\| \leqq (\mu - \rho(A))^{-1}S_p$$

and

$$\begin{aligned}\|(\mu I - A)^{-1}|(\mu I - A)|\| &\leqq \|(\mu I - A)^{-1}|(\mu I + A)|\| \\ &= \|(\mu I - A)^{-1}|(2\mu I - (\mu I - A))|\| \\ &= \|2\mu(\mu I - A)^{-1} - I\| \leqq 2\mu\|(\mu I - A)^{-1}\|.\end{aligned}$$

Thus, clearly,

$$\|(\mu I - A)^{-1}|(\mu I - A)|\| \leqq 2\mu(\mu - \rho(A))^{-1}S_p.$$

The theorem is now proved. □

Theorem 4 shows that if $4S_pS_y\hat{\mu}_s^2 u \leqq f^{-1}\delta_m(\hat{\mu}_s - \rho(A))$, where u is the computer precision, then the condition (3.2) of Theorem 3 will be satisfied for s, for which the algorithm terminates. We remark that as long as $(\hat{\mu}_s I - A)\hat{x}_s \approx \hat{y}_s$, then we have that $\hat{\mu}_s - \rho(A) \approx \|\hat{x}_s\|^{-1}$. Therefore, with $\hat{\mu}_s - \rho(A) \approx u^{1/2}\hat{\mu}_s$, one can expect that the condition (3.10) will be satisfied when $\|\hat{x}_s\| \approx u^{1/2}$ (see also [10, § 9.47]).

4. Numerical experiments. Test matrices whose Perron vectors have various oscillatory properties can be generated, for example, as follows. Take any matrix with positive entries, and divide each row by the sum of the elements of this row. This is the matrix B. Choose any n positive numbers $d_1, d_2, \ldots, d_n$, and let

$$D = \operatorname{diag}(d_1, d_2, \ldots, d_n).$$

Then

$$A = DBD^{-1} \tag{4.1}$$

is positive (and hence irreducible), the Perron root of A is equal to 1, and the Perron vector of A is $(d_1, d_2, \ldots, d_n)^T$. One possible choice of B is with all its entries equal to $1/n$ and $d_j = d^{j-1}$, $j = 1, \ldots, n$, with $d < 1$. In this case a right Perron vector is $p = (1, d, \ldots, d^{n-1})^T$ and the left Perron vector is $q = (d^{n-1}, \ldots, d, 1)^T$. It is clear that the usual condition number of the Perron root $k(F, A)$ given by (2.3) can be arbitrarily large in this case, since $1/q^T p = n^{-1} d^{1-n}$.

We have extensively tested two algorithms on such matrices, the algorithm of § 3 and the power-squaring method given by (1.6). The reason for testing the latter algorithm is its obvious numerical stability: there are no subtractions of numbers with the same sign and, with appropriate scaling, no overflows in the course of this method. The power-squaring method requires n^3 multiplication operations for each iteration step, while the algorithm we suggested in the previous section requires about $\frac{2}{3}n^3$ operations of multiplication and division per step because of the elimination involved. The experiments were performed with a computer precision $u = 2^{-52} \approx 10^{-16}$ using MATLAB on the SUN 3/50 workstation.

We observed that the algorithm of § 3 always converged quadratically to the exact Perron root, and with $\|x_s\| > 10^8 = u^{-1/2}$ the error in the computed μ_{s+1} was of $O(u) = O(10^{-16})$. The power-squaring method always converged more slowly, sometimes significantly so, than the first algorithm, which is not surprising because it has only a geometric rate of convergence. The comparison of the two methods on a parallel or vector machine could change their relative performances but we have not yet carried out such experiments.

In the following tables we give some typical numerical examples. The iteration is stopped when $\|x_s\| \geqq m_x$. In the tables, ε_1 denotes $1/\|x_s\|$ and ε_2 denotes $\operatorname{osc}(y_s/y_{s+1})$. In Table 1 we present a reproducible result, where A is the 20×20 matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \varepsilon & 0 & 0 & \cdots & 0 \end{bmatrix}$$

TABLE 1
$\varepsilon = 0.5 \times 10^{20}$.

m_x	ε_1	ε_2	$\mu_s - \rho$	$\lvert \mu_s - \varepsilon_1 - \rho \rvert$	$\mu_{s+1} - \rho$
1.2×10^4	4.1×10^{-5}	5.1×10^{-2}	4.1×10^{-5}	1.4×10^{-9}	1.0×10^{-8}
1.2×10^8	1.0×10^{-8}	1.3×10^{-7}	1.0×10^{-8}	6.8×10^{-16}	7.7×10^{-16}
1.2×10^{10}	7.7×10^{-16}	8.4×10^{-15}	7.7×10^{-16}	0	0
1.2×10^{12}	7.7×10^{-16}	8.4×10^{-15}	7.7×10^{-16}	0	0

TABLE 2

$\varepsilon = 1.2 \times 10^{-16}$.

m_x	ε_1	ε_2	$\mu_s - \rho$	$\lvert\mu_s - \varepsilon_1 - \rho\rvert$	$\mu_{s+1} - \rho$
1.2×10^{4}	1.4×10^{-4}	5.8×10^{-3}	1.4×10^{-4}	1.3×10^{-7}	4.3×10^{-7}
1.2×10^{8}	3.6×10^{-12}	1.4×10^{-10}	3.6×10^{-12}	1.38×10^{-18}	2.7×10^{-16}
1.2×10^{8}	3.6×10^{-12}	1.4×10^{-10}	3.6×10^{-12}	1.38×10^{-18}	2.7×10^{-17}
1.2×10^{12}	2.7×10^{-17}	6.6×10^{-16}	2.7×10^{-17}	6.9×10^{-18}	2.7×10^{-17}

with $\varepsilon = (0.5)^{20} = 9.5367 \times 10^{-7}$. The initial vector for our algorithm was chosen to be $y_0 = (1, 1, \ldots, 1)^T$. It is easy to see that $\rho(A) = \varepsilon^{1/20}$, $p = (1, \varepsilon^{1/n}, \ldots, \varepsilon^{(n-1)/n})^T$, $q = (\varepsilon^{(n-1)/n}, \ldots, \varepsilon^{1/n}, 1)^T$, and $q^T p = n\varepsilon^{(n-1)/n} \approx 10^{-5}$. It takes 11 steps to get the results on the first line of Table 1, 12 steps on the second line, and 13 steps on the third and fourth lines.

In Table 2 we give results for the same A above, but with $\varepsilon = (0.16)^{20} = 1.2 \times 10^{-16}$. It takes 21 steps to get the results on the first line of Table 2, 23 steps on the second and third lines, and 24 steps on the fourth line.

We see that for $s = 23$, when $m_x = u^{-1/2}$, the computed value μ_{s+1} approximates $\rho(A)$ within the computer precision. The relatively large value of $q^T p$ has no effect on the accuracy of the algorithm. We continued experiments with this matrix for decreasing $\varepsilon = 10^{-14}$, 10^{-15}, 10^{-16}, and 10^{-17}. The computed Perron root was accurate, and the results were similar, e.g., the Perron root was computed exactly after the condition $m_x^{-1} < u^{1/2}$ was satisfied. We also tested the standard QR algorithm of MATLAB (the function eig (A)) for these values of ε. For $\varepsilon = 10^{-14}$ the QR algorithm lost two significant figures, for $\varepsilon = 10^{-15}$ it lost three, and for $\varepsilon = 10^{-16}$ it lost four. For $\varepsilon = 10^{-17}$ the QR algorithm computed no significant figures at all, namely, it gave 20 complex eigenvalues $\lambda_1, \ldots, \lambda_{20}$ such that $0.06 > |\mathrm{Re}(\lambda_i)| > 0.04$ and $0.06 > |\mathrm{Im}(\lambda_i)| > 0.028$, while the value of the Perron root in this case was $10^{-17/20} = 0.141 \ldots$.

In Table 3 we present a typical example of experiments with 20×20 matrices of the form (4.1) in which B was generated randomly. The results are very similar to those of Table 1.

These examples, which are representative of a large number of other experiments that we performed, demonstrate that the separation angle between the right and left eigenvectors has no effect on the accuracy of the computed Perron root. They also show that the stopping criterion $\|x_s\|^{-1} < u^{1/2}$ gives the maximum possible accuracy and that there is no need to use a threshold smaller than $u^{1/2}$.

TABLE 3

m_x	ε_1	ε_2	$\mu_s - \rho$	$\lvert\mu_s - \varepsilon_1 - \rho\rvert$	$\mu_{s+1} - \rho$
1.2×10^{4}	1.5×10^{-5}	1.0×10^{-2}	1.5×10^{-5}	2.2×10^{-7}	4.2×10^{-10}
1.2×10^{8}	1.3×10^{-9}	8.5×10^{-5}	1.3×10^{-9}	5.3×10^{-14}	2.2×10^{-16}
1.2×10^{10}	3.3×10^{-11}	8.4×10^{-11}	3.3×10^{-11}	5.4×10^{-17}	2.2×10^{-16}
1.2×10^{12}	1.3×10^{-15}	1.8×10^{-13}	2.0×10^{-15}	2.2×10^{-16}	4.4×10^{-16}

REFERENCES

[1] M. ARIOLI, J. W. DEMMEL, AND I. S. DUFF, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.

[2] A. A. AHAC, J. J. BUONI, AND D. D. OLESKY, *Stable* LU *factorization of H-matrices*, Linear Algebra Appl., 99 (1988), pp. 97–110.

[3] A. BERMAN AND R. PLEMMONS, *Nonnegative Matrices in Mathematical Sciences*, Academic Press, New York, 1979.

[4] L. ELSNER, *Inverse iteration for calculating the spectral radius of a nonnegative irreducible matrix*, Linear Algebra Appl., 15 (1976), pp. 235–242.

[5] S. FRIEDLAND, *Revisiting matrix squaring*, Linear Algebra Appl., 154/156 (1991), pp. 59–63.

[6] I. GOHBERG AND I. KOLTRACHT, *Componentwise mixed and structured condition numbers*, submitted.

[7] T. NODA, *Note on the computation of the maximal eigenvalue of a nonnegative irreducible matrix*, Numer. Math., 17 (1971), pp. 382–386.

[8] E. SENETA, *Non-negative Matrices and Markov Chains*, Springer-Verlag, New York, 1981.

[9] R. D. SKEEL, *Scaling for numerical stability in Gaussian elimination*, J. Assoc. Comput. Mach., 26 (1979), pp. 494–526.

[10] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, U.K., 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 468–493, April 1993

011

THE USE OF PIVOTING TO IMPROVE THE NUMERICAL PERFORMANCE OF ALGORITHMS FOR TOEPLITZ MATRICES*

DOUGLAS SWEET†

Abstract. Classical $O(n^2)$ Toeplitz solvers break down if any leading principal submatrices are singular, and as might be expected, the occurrence of nearly singular leading submatrices can cause a serious loss of accuracy in these solvers. In this paper, a pivoting scheme has been incorporated into the Toeplitz solver of Bareiss which allows near-singularities to be treated without significant loss of accuracy. As special cases, the pivoted Toeplitz solver also handles exact singularities and numerical singularities—the latter appear as near-singularities when finite-precision floating-point arithmetic is used. The Bareiss algorithm can also be used to compute the LU factors of a Toeplitz matrix, and it is shown that a restricted version of the pivoted Toeplitz solver produces an LU factorization of a row- and column-permuted version of the original Toeplitz matrix. Finally, it is shown how the inverter of Trench and Zohar can be derived from the multipliers produced by the Bareiss algorithm, and this connection is used to derive a pivoted version of the Trench–Zohar algorithm from the pivoted Bareiss algorithm.

Key words. Toeplitz matrices, pivoting, error analysis

AMS(MOS) subject classification. 65F30

1. Introduction.

A Toeplitz matrix is one whose entries are constant along each diagonal, i.e., if T is Toeplitz, $t_{ij} = t_{i-1,j-1}$ for $2 \leqq i, j \leqq n$, the order of T. Positive-definite Toeplitz matrices have many applications in signal processing, statistics, and other fields [3], [11]. For example, the autocorrelation matrix of a stationary stochastic process is Toeplitz, as is the cross-correlation matrix of a linear, equally spaced array in a noise field with distant, uncorrelated sources. There are several "classical" algorithms that solve Toeplitz systems in $O(n^2)$ operations. These include the algorithms of Levinson [9], Trench [14], and Bareiss [2], to name a few of the more well known methods. It has been shown [3], [4], [11] that these classical algorithms are stable in the symmetric positive-definite case.

Recently, attention has been turned to arbitrary Toeplitz matrices, where the assumption of positive-definiteness has been dropped. The computation of Padé approximants [1] can require the solution of a general Toeplitz system. The design of eigenfilters, investigated by Makhoul [10], Kung [8], and others, may require the solution of an indefinite Toeplitz system. The classical methods still apply to indefinite systems, provided that all the leading principal submatrices are nonsingular. If this property is not satisfied, these methods fail. However, recently Delsarte, Genin, and Kamp [5], Heinig [6], and others have proposed methods to generalize the Levinson algorithm to solve systems with singular leading submatrices.

It would be expected that if the normal Levinson algorithm breaks down when a leading submatrix is singular, a serious loss of accuracy could occur using finite precision arithmetic when a leading submatrix is nearly singular. This has been shown by Bunch [3] for the Trench algorithm and for classical $O(n^2)$ Toeplitz solvers related to the Trench algorithm. In § 2, we analyze one of these solvers, the Bareiss algorithm [2], in more detail; we show that nearly singular leading submatrices (NSLSs) cause small "pivot" values that result in large multipliers, which in turn propagate large increases in relative error. We then see that there is some freedom in the Bareiss algorithm to select alternative pivots, and such a pivoting technique is proposed in § 3 to handle cases of NSLSs without

* Received by the editors November 15, 1989; accepted for publication May 22, 1991.

† Electronics Research Laboratory, Defence Science and Technology Organisation, Salisbury SA 5108, Australia (dougs@ewd.dsto.gov.au). This work was completed while the author was with the Department of Computer Science, James Cook University of North Queensland, Townsville, Qld. 4811, Australia.

significant loss of accuracy. In his original paper, Bareiss proposed a pivoting scheme to handle exactly singular leading submatrices; however, the present method, which applies to singular or nearly singular leading submatrices, is different from that of Bareiss, and of more general application.

The Bareiss algorithm can also be used to compute the LU factors of a Toeplitz matrix, and it is shown in § 4 how a restricted version of the pivoted Bareiss algorithm (PBA) can be used to compute an LU factorization of a row- and column-permuted version of the original Toeplitz matrix. In § 5, it is shown how Trench's Toeplitz inverter can be derived from the multipliers produced by the Bareiss algorithm, and hence a pivoted Trench algorithm is derived from PBA. Numerical examples are given demonstrating the improved performance of all the pivoted algorithms. Finally, some conclusions are drawn, and pointers to further work indicated.

A preliminary version of this paper was presented at the SPIE 1986 Conference on Advanced Algorithms and Architectures for Signal Processing [12]. The material has been greatly expanded and modified in this paper, with the inclusion of full descriptions of the pivoting algorithms with diagrams, proofs and major results, a new and faster algorithm for "Backtrack Procedure C" (see § 3.6), and a more general method for extracting the upper triangle for back-substitution. In certain cases, the algorithm of [12] does not terminate normally, and an *augmentation* procedure is described in § 3.9 to handle these cases.

Remarks on notation. The symbol ":=" in an algorithm denotes assignment. For any vector, $\mathbf{a}$, $\mathbf{a}^R$, and $\mathbf{a}^{RT}$ denote the reversal of $\mathbf{a}$ and $\mathbf{a}^T$, respectively, and $\mathbf{a}_{i:j}$ denotes $[a_i, \dots, a_j]^T$. For any matrix A, $\mathbf{a}_{i\cdot}$ and $\mathbf{a}_{\cdot j}$ denote row i and column j, respectively, of A; $\mathbf{a}_{i,j:k}$ denotes $[a_{ij}, \dots, a_{ik}]$; A_j denotes the leading $j \times j$ submatrix of A; and $A_{k:l\cdot}$ denotes rows k to l to A.

2. Error analysis of the Bareiss algorithm.

2. Error analysis of the Bareiss algorithm. Here we present the Bareiss algorithm and demonstrate with a numerical example the loss of accuracy that can occur when a leading submatrix is nearly singular. We show by means of a backward error analysis how this loss of accuracy occurs.

2.1. The Bareiss algorithm. The Bareiss algorithm solves the general Toeplitz system

$$T\mathbf{x} = \mathbf{b} \tag{2.1}$$

by transforming (2.1) successively into the systems $T^{(-0)}\mathbf{x} = \mathbf{b}^{(-0)}$, $T^{(+0)}\mathbf{x} = \mathbf{b}^{(+0)}$, $T^{(-1)}\mathbf{x} = \mathbf{b}^{(-1)}$, $T^{(+1)}\mathbf{x} = \mathbf{b}^{(+1)}, \dots, T^{(1-n)}\mathbf{x} = \mathbf{b}^{(1-n)}$, $T^{(n-1)}\mathbf{x} = \mathbf{b}^{(n-1)}$, where $T^{(-j)}$ has zeros along the j subdiagonals, and $T^{(j)}$ has zeros along the j superdiagonals. Define the shift matrices $Z_{\pm j}$ by

$$(Z_{\pm j})_{kl} = \delta_{k\pm j,l}, \tag{2.2}$$

where δ is the discrete δ-function. Premultiplication by $Z_{\pm j}$ has the effect of moving a matrix up (down) by j rows and zeroing the bottom (top) j rows of the result. The Bareiss algorithm is as follows.

$$\begin{aligned} &T^{(-0)} := T^{(0)} := T; \qquad \mathbf{b}^{(-0)} := \mathbf{b}^{(0)} := \mathbf{b}. \\ &\textbf{For } j := 1 \textbf{ to } n-1 \textbf{ do:} \end{aligned} \tag{2.3}$$

$$m_{-j} := t_{j+1,1}^{(1-j)} / t_{11}^{(j-1)}, \tag{2.4}$$

$$T^{(-j)} := T^{(1-j)} - m_{-j} Z_{-j} T^{(j-1)}, \qquad \mathbf{b}^{(-j)} := \mathbf{b}^{(1-j)} - m_{-j} Z_{-j} \mathbf{b}^{(j-1)}, \tag{2.5}$$

$$m_j := t_{n-j,n}^{(j-1)} / t_{nn}^{(-j)}, \tag{2.6}$$

$$T^{(j)} := T^{(j-1)} - m_j Z_j T^{(-j)}, \qquad \mathbf{b}^{(j)} := \mathbf{b}^{(j-1)} - m_j Z_j \mathbf{b}^{(-j)}. \tag{2.7}$$

It is easily shown that after step j, $T^{(-j)}$ and $T^{(j)}$ have the form shown in Fig. 2.1. Since the nonnull areas of $Z_{-j}T^{(j-1)}$ and $Z_jT^{(-j)}$ are Toeplitz, there are only $O(n)$ multiplications per Bareiss step, and the complete reduction requires $O(n^2)$ multiplications.

We now introduce an alternative notation which will be useful in the sequel. Let $t_A^{(-j)}$ and $t_A^{(+j)}$ denote the elements in the first Toeplitz nonzero diagonal *above* the zero-band (ZB) (of j diagonals) of $T^{(-j)}$ and $T^{(+j)}$, respectively (see Fig. 2.1) and let $t_B^{(-j)}$ and $t_B^{(+j)}$ denote the elements in the first Toeplitz nonzero diagonal *below* the ZB of $T^{(-j)}$ and $T^{(+j)}$, respectively. We see from Fig. 2.1 that

$$t_A^{(-j)} = t_{nn}^{(-j)}, \qquad t_B^{(-j)} = t_{j+2,1}^{(-j)}, \tag{2.8a}$$

$$t_A^{(j)} = t_{n-j-1,n}^{(j)} \quad \text{and} \quad t_B^{(j)} = t_{11}^{(j)}. \tag{2.8b}$$

Let $T_*^{(\pm j)}$, $T_{*Fk}^{(\pm j)}$, and $T_{*Lk}^{(\pm j)}$ denote, respectively, the *Toeplitz part*, the Toeplitz part *excluding the first k rows*, and the Toeplitz part *excluding the last k rows* of $T^{(\pm j)}$; and let $\mathbf{b}_*^{(\pm j)}$, $\mathbf{b}_{*Fk}^{(\pm j)}$, and $\mathbf{b}_{*Lk}^{(\pm j)}$ denote, respectively, the corresponding elements of $\mathbf{b}^{(\pm j)}$. From Fig. 2.1 it can be seen that for the Bareiss algorithm, $T_*^{(-j)} = T_{j+1:n\cdot}^{(-j)}$ and $T_*^{(j)} = T_{1:n-j\cdot}^{(j)}$. Using this notation and (2.8), (2.4)–(2.7) can be written as

$$m_{-j} := t_B^{(1-j)}/t_B^{(j-1)}, \tag{2.9}$$

$$T_*^{(-j)} := T_{*F1}^{(1-j)} - m_{-j}T_{*L1}^{(j-1)}, \qquad \mathbf{b}_*^{(-j)} := \mathbf{b}_{*F1}^{(1-j)} - m_{-j}\mathbf{b}_{*L1}^{(j-1)}, \tag{2.10}$$

$$m_j := t_A^{(j-1)}/t_A^{(-j)}, \tag{2.11}$$

$$T_*^{(j)} := T_{*L1}^{(j-1)} - m_jT_*^{(-j)}, \qquad \mathbf{b}_*^{(j)} := \mathbf{b}_{*L1}^{(j-1)} - m_j\mathbf{b}_*^{(-j)}. \tag{2.12}$$

2.2. Loss of accuracy in the Bareiss algorithm. The system $T\mathbf{x} = \mathbf{b}$, with

$$\mathbf{t}_{1\cdot} = [4, 8, 1, 6, 2, 3], \tag{2.13a}$$

$$\mathbf{t}_{\cdot 1} = [4, 6, (71/15) + a, 5, 3, 1]^T, \tag{2.13b}$$

$$\mathbf{b} = T[1, \dots, 1]^T, \tag{2.13c}$$

where $a = 5 \times 10^{-8}$, was solved in double precision (56-bit mantissa) using the Bareiss algorithm. The error vector $\bar{\mathbf{x}} - \mathbf{x}$ had infinity norm 6.9×10^{-2}. Now T_3, the 3×3 leading submatrix of T, would be exactly singular if $a = 0$. The near-singularity of T_3 caused a serious loss of accuracy, even though T was well conditioned. This result is in line with Bunch's general analysis of the stability of $O(n^2)$ Toeplitz algorithms. We next examine error propagation in the Bareiss algorithm in more detail, and conclude that as for unpivoted Gaussian elimination, the loss of accuracy is due to the occurrence of small pivots which produce large multipliers $m_{\pm j}$ in (2.9) and (2.11).

2.3. Bareiss pivots. Consider the form of the iterates (Fig. 2.1) produced by the Bareiss recursion (2.9)–(2.12). We can see that in step $(-j)$, the diagonal containing

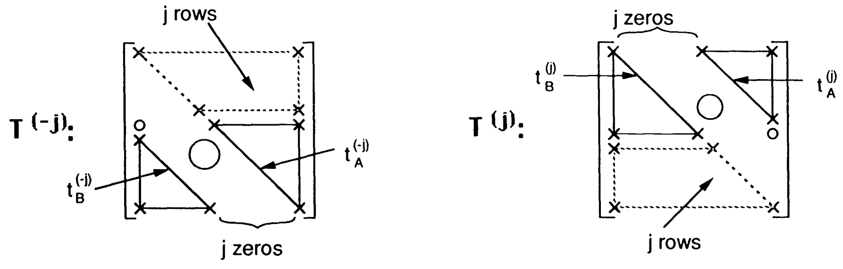


FIG. 2.1. *Form of $T^{(-j)}$ and $T^{(j)}$. Areas bounded by —— have Toeplitz form, areas bounded by - - - have general form, and remaining areas are null.*

$t_B^{(j-1)}$ is used to eliminate the diagonal containing $t_B^{(1-j)}$. Thus $t_B^{(j-1)}$ is the *pivot* element for step $(-j)$. Similarly, $t_A^{(-j)}$ is the pivot element for step j. It can be shown by straightforward algebraic manipulation (see Appendix A) that for the case

$$\|T_{j+1}^{-1}\| > \|T_j^{-1}\|, \tag{2.14a}$$

the following applies:

$$|t_A^{(-j)}| \leqq \frac{\|T_{j+1}\|(1+\beta\kappa(T_j))^2}{\kappa(T_{j+1})-\beta\kappa(T_j)}, \tag{2.14b}$$

where $\|\cdot\|$ is any matrix norm $\beta = \|T_{j+1}\|/\|T_j\|$, and $\kappa(T_i)$ is the condition number of T_i, defined by $\|T_i\|\|T_i^{-1}\|$. Note that because of (2.14a), the denominator in (2.14b) is positive. Equation (2.14b) shows that when T_j is well conditioned but T_{j+1} is not, the pivot $t_A^{(-j)}$ will be small, and hence, by (2.11), will produce a large multiplier. As we see below, large multipliers can cause large relative errors in the Bareiss algorithm.

2.4. Backward error analysis of the Bareiss algorithm. Recall that the aim of the Bareiss algorithm is to reduce the system (2.1) to the upper and lower triangular systems

$$T^{(1-n)}\mathbf{x} = \mathbf{b}^{(1-n)}, \tag{2.15a}$$

$$T^{(n-1)}\mathbf{x} = \mathbf{b}^{(n-1)}. \tag{2.15b}$$

Let bars denote computed quantities, and let $\mathbf{x}_u$ and $\mathbf{x}_l$ be the solutions, respectively, of

$$\bar{T}^{(1-n)}\mathbf{x}_u = \bar{\mathbf{b}}^{(1-n)}, \tag{2.16a}$$

$$\bar{T}^{(n-1)}\mathbf{x}_l = \bar{\mathbf{b}}^{(n-1)}. \tag{2.16b}$$

Then it can be shown after considerable manipulation (see Appendix B) that $\mathbf{x}_u$ and $\mathbf{x}_l$ are, to a first-order approximation, the solutions, respectively, of the perturbed systems

$$[T+LF]\mathbf{x}_u = \mathbf{b} + L\mathbf{d}, \tag{2.17}$$

$$[T+U^{T2}R/t_{11}]\mathbf{x}_l = \mathbf{b} + U^{T2}\mathbf{g}/t_{11}. \tag{2.18}$$

Here L is the unit lower-triangular factor of T, U is the upper-triangular factor, $T2$ denotes transpose about the secondary diagonal, and the elements of the perturbation matrices and vectors are bounded by

$$|f_{kl}|, \quad |r_{kl}| \leqq 4(n-1)\mu\alpha\tau, \tag{2.19a, b}$$

$$|d_k|, \quad |g_k| \leqq 4(n-1)\mu\alpha\beta, \tag{2.20a, b}$$

where μ is the machine precision, $\tau = \max_{k,l}|t_{kl}|$, $\beta = \max_k |b_k|$, and

$$\alpha = \prod_{l=1-n}^{n-1}(1+|m_l|). \tag{2.21}$$

Equations (2.17)–(2.21) show that if there are any large multipliers, such as would occur from small pivots resulting from NSLSs, the error in the computed solution can be large.

3. The pivoted Bareiss algorithm. Here we first outline the basic pivoting procedure (BPP). It will be shown that BPP can be used to handle cases of isolated NSLSs. Then it will be indicated how to combine several steps of BPP with backtracking to produce a composite pivoting procedure (CPP) that handles several cases of contiguous NSLSs.

In all of what follows, only the operations on the iterates $T^{(\pm j)}$ will be described. It is understood that corresponding operations will also be performed on the right-hand sides $\mathbf{b}^{(\pm j)}$.

3.1. The basic pivoting procedure. We noted above that in step $-j$ of the Bareiss algorithm, $t_B^{(j-1)}$ is used as pivot to eliminate $t_B^{(1-j)}$. We could instead use $t_A^{(j-1)}$ to eliminate $t_A^{(1-j)}$. Thus there is a two-way choice of pivot at step $(-j)$.

At step j in the normal Bareiss algorithm (in which $T^{(j)}$ is computed), $t_A^{(-j)}$ is used as pivot to eliminate $t_A^{(j-1)}$. We could instead use $t_B^{(-j)}$ to eliminate $t_B^{(j-1)}$. Thus there is also a choice at step j. As well as having a two-way choice at each step, we also have the choice of whether to perform step $-j$ or step j first. These choices motivate the BPP.

3.2. Use of BPP to handle isolated NSLS. Equation (2.14b) shows that if T_{j+1} is nearly singular, $t_A^{(-j)}$ is small (zero if T_{j+1} is singular). Hence the normal Bareiss algorithm can suffer a large loss of accuracy at step j in this case. However, if T_{j+2} is well conditioned, $t_B^{(-j)}$ will not be small, as is indicated by the following lemma.

LEMMA 3.1. *Let* $t_A^{(-j)}$ *and* $t_B^{(-j)}$ *be as defined above. Then*

$$|t_A^{(-j)}| + |t_B^{(-j)}| \geqq \|T_{j+2}\|/\kappa(T_{j+2}). \tag{3.1}$$

The proof of this lemma is given in Appendix C. Because of (3.1), we can use BPP as follows to handle the case where T_{j+1}, but not T_{j+2}, is ill conditioned.

ALGORITHM 3.1.

1. Run the normal Bareiss algorithm to step $(-j)$. Then $t_A^{(-j)}$, but not $t_B^{(-j)}$, will be small.
2. With $t_B^{(-j)}$ as pivot, eliminate $t_B^{(j-1)}$ in step j:
$$m_j := t_B^{(j-1)}/t_B^{(-j)}, \qquad T_*^{(j)} := T_{*L2}^{(j-1)} - m_j T_{*F1}^{(-j)}.$$
3. With $t_A^{(j)}$ as pivot, eliminate $t_A^{(-j)}$ in step $(-j-1)$:
$$m_{-j-1} := t_A^{(-j)}/t_A^{(j)}, \qquad T_*^{(-j-1)} := T_{*L1}^{(-j)} - m_{-j-1} T_*^{(j)}.$$
4. With $t_A^{(-j-1)}$ as pivot, eliminate $t_A^{(j)}$ in step $j+1$:
$$m_{j+1} := t_A^{(j)}/t_A^{(-j-1)}, \qquad T_*^{(j+1)} := T_{*F1}^{(j)} - m_{j+1} T_{*L1}^{(-j-1)}.$$

   It can be verified that at this juncture $T^{(-j-1)}$ and $T^{(j+1)}$ have the forms shown in Fig. 3.1. By appropriately permuting the rows between $T^{(-j-1)}$ and $T^{(j+1)}$, these can be restored to "Bareiss form" (Fig. 2.1), apart from the nonzero elements "a" and "b," which after permutation, appear in $T^{(j+1)}$. The reduction can then be completed as follows:
5. Continue with the Bareiss algorithm, repeating steps 2–4 if necessary. At the end of the reduction, $T^{(1-n)}$ (but not $T^{(n-1)}$) will be upper triangular. Solve the upper-triangular system by back-substitution.

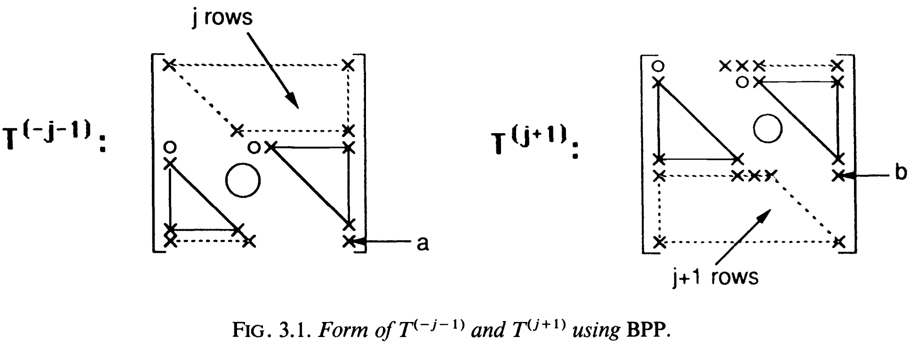


FIG. 3.1. *Form of* $T^{(-j-1)}$ *and* $T^{(j+1)}$ *using* BPP.

The pivot threshold. We have not yet addressed the question of how "small" $t_A^{(-j)}$ should be before pivoting is required. The following heuristic should handle all cases of isolated NSLSs without causing unnecessary pivoting:

(i) find $t_{\max}^{(-j)}$, the element in the Toeplitz part of $T^{(-j)}$ with largest modulus;

(ii) carry out the above pivoting procedure if $|t_A^{(-j)}| < 0.001 |t_{\max}^{(-j)}|$ and $|t_A^{(-j)}| < |t_B^{(-j)}|$.

A higher threshold, say $|t_A^{(-j)}| < 0.1 |t_{\max}^{(-j)}|$, should handle even cases of mildly ill conditioned isolated leading submatrices, though at the expense of extra pivoting. The latter strategy is analogous to that used in sparse mixing pivoting, where a similar threshold is used with a view to pivoting only when necessary to avoid a serious loss of accuracy, so that "fill-in" is minimized.

Numerical results. The above procedure was run on the system $T\mathbf{x} = \mathbf{b}$ specified by (2.13). A value of 1.1×10^{-15} was obtained for $\|\bar{\mathbf{x}} - \mathbf{x}\|_\infty$, compared to 6.9×10^{-2} in the unpivoted case (see § 2.2).

3.3. Composite pivoting procedure. In the following, we let $t_k^{(\pm j)}$ and $t_{-k}^{(\pm j)}$ denote the Toeplitz diagonals k places, respectively, above and below the ZBs of $T^{(\pm j)}$. Note that $t_1^{(\pm j)} = t_A^{(\pm j)}$ and $t_{-1}^{(\pm j)} = t_B^{(\pm j)}$.

There are some cases in which both $t_A^{(-j)}$ and $t_B^{(-j)}$, as well as other Toeplitz diagonals adjoining these, are small, and the above procedure will not produce satisfactory results. This implies that T_j *and* one or more adjacent higher-order leading submatrices of T are near-singular, for by a straightforward generalization of the proof of Lemma 3.1, it can be shown that

$$\sum_{\substack{k=-s \\ k \neq 0}}^{r} |t_k^{(-j)}| \geqq \|T_{j+r+s}\| / \kappa(T_{j+r+s}).$$

Thus, if $|t_k^{(-j)}|$, $k = -s, \ldots, -1, 1, \ldots, r$ are all small, then $\kappa(T_{j+1}), \ldots, \kappa(T_{j+r+s})$ are all large.

If $t_A^{(-j)}$ and $t_B^{(-j)}$ are both unsatisfactory as pivots, it would be desirable to use some other, larger diagonal $t_{-k}^{(-j)}$ or $t_l^{(-j)}$ as pivot. This cannot be done immediately, as several of the zero diagonals in $T^{(j-1)}$ will be destroyed. However, if by some means the Toeplitz ZB can be "moved" up or down so that a satisfactory pivot is made available (the selection of a suitable pivot is discussed later), the small diagonals, which reappear below the ZB if the latter was moved up (and above the ZB if the latter was moved down), can be eliminated by a sequence of ordinary Bareiss cycles.

Figure 3.2(a) shows the ZB in its normal position as given by the Bareiss algorithm, and Fig. 3.2(b) shows the ZB moved up so that it adjoins the desired pivot.

Figure 3.3(a) shows the situation when the desired pivot is *below* the ZB. Note that p is negative here. Figures 3.3(b) and 3.3(c) show the ZB moved down $|p| - 1$ and $|p|$ places, respectively. Observe that in Fig. 3.3(c), the desired pivot has appeared *above* the ZB. This is because whenever the ZB is moved, the diagonals displaced by the shift appear on the *other* side of the ZB.

In the procedures to be outlined, we always move the ZB so that the desired pivot appears just *above* the ZB (rather than just below). This enables us to continue the reduction using normal Bareiss steps, starting with the new pivot $\tilde{t}_1^{(-j)}$.

We now describe three alternative methods to move the ZB. All involve *backtracking* one or more Bareiss cycles, i.e., recovering iterates $T^{\pm(j-|p|)}$, $|p| \geqq 1$, then carrying out p cycles of BPP in such a manner as to move the ZB in the desired direction, so that it adjoins a large pivot; thus these methods are called "backtrack procedures." Which procedure to select in a given situation is discussed below.

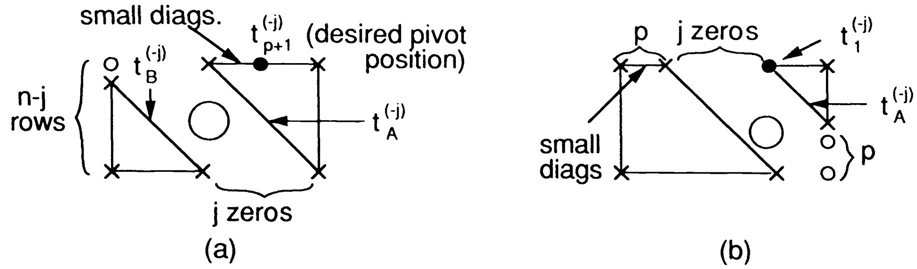


FIG. 3.2. $T_*^{(-j)}$: (a) *normal Bareiss form*, (b) *with* ZB *moved up p places to adjoin* $t_{p+1}^{(-j)}$. ● *indicates the position of the desired pivot.*

One might ask whether "moving" the ZB changes the desired pivot significantly, so that it is no longer sufficiently large. The next lemma shows that if, say, $t_1^{(-j)}$ $(=t_A^{(-j)})$, $t_2^{(-j)}, \dots, t_p^{(-j)}$ are all small, but $t_{p+1}^{(-j)}$ is not, then moving the ZB up p places by any procedure to make $t_{p+1}^{(-j)}$ available for pivoting will not greatly change the latter. A similar remark can be made when the ZB is moved down.

LEMMA 3.2. *Let $T_{j;p}$ be the jth Toeplitz submatrix of T displaced p places right of* (*below*) *the main diagonal for p positive* (*negative*), *and assume that $T_{j;p}$ is nonsingular. For the Toeplitz part of $T^{(-j)}$, let $t_{\pm k}^{(-j)}$ be as defined above, and let $\tilde{T}^{(-j)}$ be the result of any pivoting procedure that moves the* ZB *to displacement p. Let k' be such that $t_{k'}^{(-j)} = t_{\max}^{(-j)}$, as defined previously. Then*

$$\frac{\tilde{t}_{k-p}^{(-j)}}{\tilde{t}_{k'-p}^{(-j)}} = \frac{\dot{t}_k^{(-j)} + \varepsilon_k}{t_{k'}^{(-j)} + \varepsilon_{k'}}, \tag{3.2}$$

where

$$\dot{t}_k^{(-j)} = \begin{cases} t_k^{(-j)}, & k = j+1-n, \dots, u-1, \quad v+1, \dots, n-j, \\ 0, & k = 0, \dots, p - \operatorname{sgn} p, \end{cases}$$

$$u = \min(0, p); \qquad v = \max(0, p),$$

$$|\varepsilon_k| \leqq \begin{cases} \beta_p \kappa(T_{j;p}) \|\mathbf{0}^T, t_1^{(-j)}, \dots, t_p^{(-j)}\|, & p > 0, \\ \beta_p \kappa(T_{j;p}) \|t_p^{(-j)}, \dots, t_{-1}^{(-j)}, \mathbf{0}^T\|, & p < 0, \end{cases} \tag{3.3}$$

and

$$\beta_p = \|T\| / \|T_{j;p}\|.$$

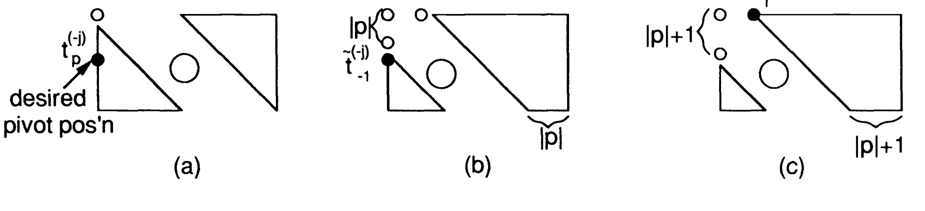


FIG. 3.3. $T_*^{(-j)}$: (a) *normal Bareiss Form*, (b) *with* ZB *moved down* $|p-1|$ *places so that the desired pivot is just below the* ZB, (c) *with* ZB *moved down* $|p|$ *places so that the desired pivot is just above the ZB.* ● *indicates the position of the desired pivot* $t_p^{(-j)}$.

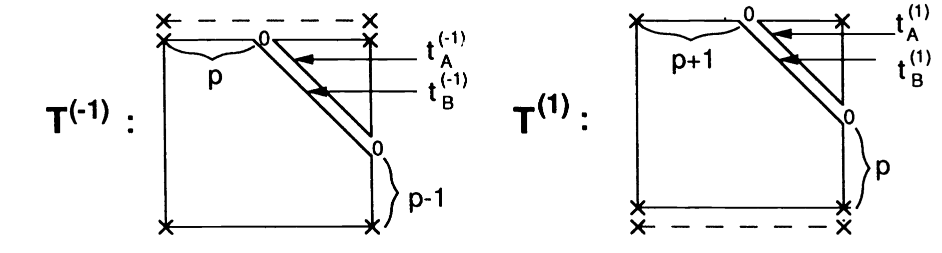


FIG. 3.4. *Form of* $T^{(-1)}$ *and* $T^{(1)}$ *after one cycle of* BPA.

The proof of this lemma is straightforward but tedious. Details are given in Appendix D. Equations (3.2) and (3.3) clearly show that if $t_{p+1}^{(-j)}$ is not small, but the diagonals between it and the ZB are, then there will not be a great relative change in the former.

We now describe each backtrack procedure in detail.

3.4. Backtrack procedure A (BPA). This procedure is only preferable if $j < |p|$, where p is the desired ZB displacement, as we must recover $T^{(0)}$ and repeat cycles 1 to j; here we eliminate, for each $T^{(-i)}$ and $T^{(i)}$, the diagonals, respectively, below and above $t_{1,1+p}$ (for $p > 0$) instead of t_{11}. For this case (ZB upshift) we proceed as follows.

1. With $t_{1,1+p}^{(+0)}$ as pivot, eliminate the diagonal containing $t_{2,1+p}^{(-0)}$ (step -1):

$$m_{-1} := t_{2,1+p}^{(-0)}/t_{1,1+p}^{(+0)}, \qquad T_*^{(-1)} := T_{*F1}^{(-0)} - m_{-1}T_{*L1}^{(+0)}.$$

2. With the diagonal containing $t_{n-p,n}^{(-1)}$ as pivot, eliminate the diagonal containing $t_{n-p-1,n}^{(+0)}$ (step 1):

$$m_1 := t_{n-p-1,n}^{(0)}/t_{n-p,n}^{(-1)}, \qquad T_*^{(1)} := T_{*L1}^{(0)} - m_1T_*^{(-1)}.$$

After this step, $T^{(-1)}$ and $T^{(1)}$ have the form shown in Fig. 3.4.

3. For $i = 2$ to j, proceed as in the Bareiss algorithm, using $t_B^{(i-1)}$ to eliminate $t_B^{(1-i)}$ (step $-i$), then $t_A^{(-i)}$ to eliminate $t_A^{(i-1)}$ (step i). The matrix operations are as in (2.9)-(2.12), with the new $t_A^{(\pm i)}$ and $t_B^{(\pm i)}$.

It is easy to see that after this procedure, $T^{(-j)}$ and $T^{(j)}$ have their ZBs displaced up p places, as illustrated in Fig. 3.5. The procedure for $p < 0$ is similar.

3.5. Backtrack procedure B (BPB). In this case, we go back to step $(-j + |p|)$ and use the basic pivoting procedure p times to move the ZB p places in the desired direction. At steps $-i$ and i, we move the ZB up one place as follows.

U-cycle. (i) With $t_A^{(i-1)}$ as pivot, eliminate $t_A^{(1-i)}$ (step $-i$):

$$m_{-i} := t_A^{(1-i)}/t_A^{(i-1)}, \qquad T_*^{(-i)} := T_{*F1}^{(1-i)} - m_{-i}T_{*L1}^{(i-1)}.$$

(ii) With $t_A^{(-i)}$ as pivot, eliminate $t_A^{(i-1)}$ (step i):

$$m_i := t_A^{(i-1)}/t_A^{(-i)}, \qquad T_*^{(i)} := T_{*F1}^{(i-1)} - m_iT_*^{(-i)}.$$

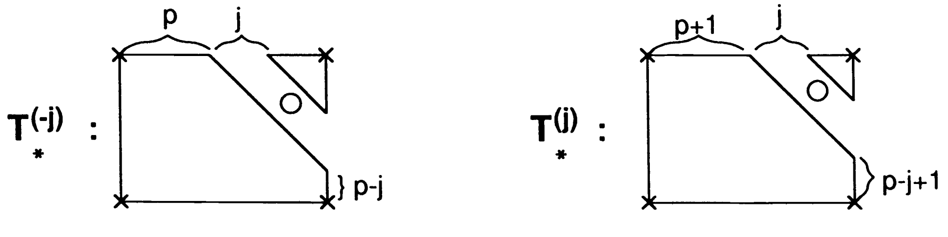


FIG. 3.5. *Form of* $T_*^{(-j)}$ *and* $T_*^{(j)}$ *after j cycles of* BPA.

Figure 3.6 shows a typical U-cycle. In $T_*^{(1-i)}$ and $T_*^{(i-1)}$, the ZB is "displaced" r places up from its "normal" Bareiss position, that is, its *zero-band displacement* (ZBD) is r. In $T_*^{(-i)}$ and $T_*^{(i)}$, ZBD is clearly $r+1$. Similarly, we move the ZB *down* one place at steps i and $-i$ as follows.

D-cycle. (i) With $t_B^{(1-i)}$ as pivot, eliminate $t_B^{(i-1)}$ (step i):

$$m_i := t_B^{(i-1)}/t_B^{(1-i)}, \qquad T_*^{(i)} := T_{*L1}^{(i-1)} - m_i T_{*F1}^{(1-i)}.$$

(ii) With $t_B^{(i)}$ as pivot, eliminate $t_B^{(1-i)}$ (step $-i$):

$$m_{-i} := t_B^{(1-i)}/t_B^{(i)}, \qquad T_*^{(-i)} := T_{*L1}^{(1-i)} - m_{-i} T_*^{(i)}.$$

Figure 3.7 shows a typical D-cycle. ZBD is clearly $-r$ in $T_*^{(1-i)}$ and $T_*^{(i-1)}$, and $-r-1$ in $T_*^{(-i)}$ and $T_*^{(i)}$. Note that in a D-cycle, step i is performed before step $-i$.

Note. In [11], U- and D-cycles are called A- and C-cycles, respectively, but the U, D notation is more descriptive for its effect on the ZB. BPB is more efficient than BPA if $|p| \leqq j$.

3.6. Backtrack procedure C (BPC). This is an alternative method to BPB for moving the ZB up or down. It is slower than BPB, but allows the pivots to be monitored continuously while the shifting is being carried out so that the shifting can continue until a satisfactory pivot is obtained. In each major step of this procedure we start with $T_*^{(j-2)}$ and $T_*^{(j-1)}$ and the corresponding *multiplier matrices* $M_*^{(j-2)}$ and $M_*^{(j-1)}$, where $M^{(j-2)}$ and $M^{(j-1)}$ are defined by

$$M^{(\pm k)}T \equiv T^{(\pm k)} \quad \text{for } k = 0, \dots, \tag{3.4}$$

and compute $\hat{T}_*^{\pm(j-2)}$, $\hat{M}_*^{\pm(j-2)}$, $\hat{T}_*^{\pm(j-1)}$, and $\hat{M}_*^{\pm(j-1)}$, where for any k, $\hat{T}^{(\pm k)}$, $\hat{M}^{(\pm k)}$ are $T^{(\pm k)}$, $M^{(\pm k)}$, respectively, with the ZB shifted one place. The following properties of the T's and M's are easily shown.

A. The $M^{(\pm k)}$ may be computed by replacing the T's and M's in the Bareiss recursion, with the initial values $M^{(\pm 0)} = I$. Thus, for each Bareiss step, we have

$$M_*^{(-j)} := M_{*F1}^{(1-j)} - m_{-j} M_{*L1}^{(j-1)}, \tag{3.5}$$

$$M_*^{(j)} := M_{*L1}^{(j-1)} - m_j M_*^{(-j)}. \tag{3.6}$$

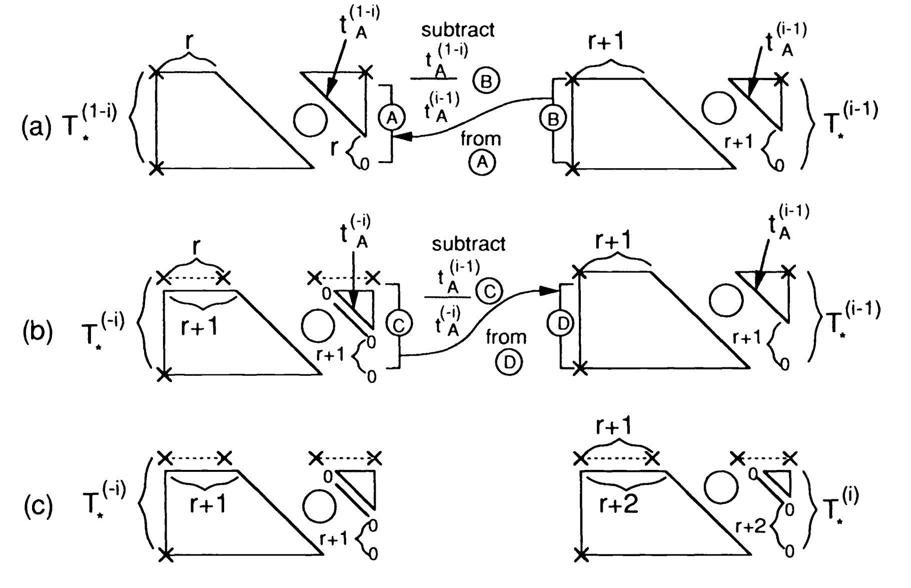


FIG. 3.6 (a) $T_*^{(1-i)}$ *and* $T_*^{(i-1)}$ *with* ZBD $= r$. (b) *After step* $-i$ *of a* U-*cycle*. (c) *After step* i *of a* U-*cycle*.

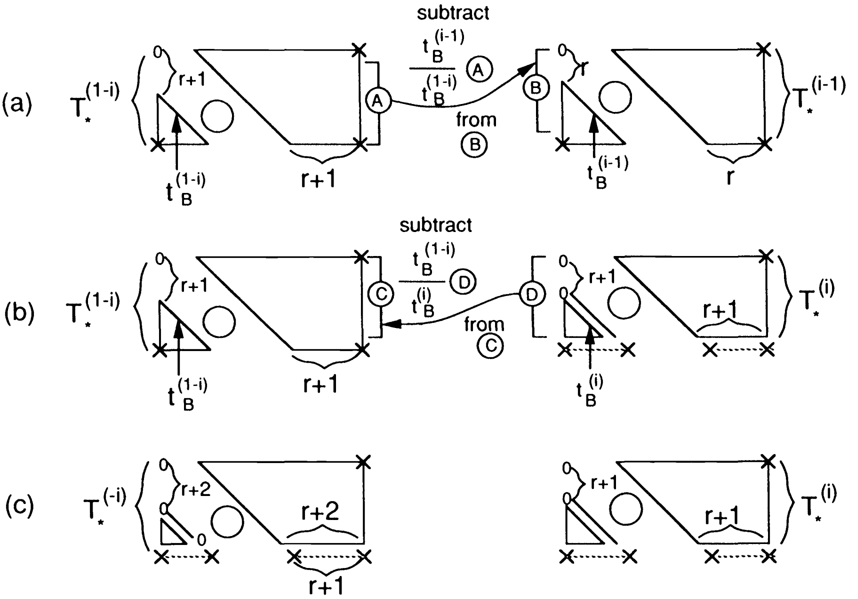


FIG. 3.7. (a) $T_*^{(1-i)}$ *and* $T_*^{(i-1)}$ *with* ZBD = $-r$. (b) *After step i of a* D*-cycle*. (c) *After step* $-i$ *of a* D*-cycle*.

B. The $M^{(\pm k)}$ have *k non-Toeplitz* and $n - k$ *Toeplitz* rows, corresponding to the non-Toeplitz and Toeplitz rows, respectively, of the $T^{(\pm k)}$. $M_*^{(-k)}$ and $M_*^{(k)}$ both have the form shown in Fig. 3.8.

C. Let $m_*^{(\pm k)}$ denote the upper nonzero diagonals of the $M_*^{(\pm k)}$ (see Fig. 3.8). We then have, for an *upward* shift in ZB,

$$\frac{\hat{T}_*^{(-k)}}{\hat{m}_*^{(-k)}} = \frac{T_*^{(k)}}{m_*^{(k)}}, \qquad k = j-2, j-1. \tag{3.7}$$

Thus $T_*^{(k)}$ is a scaled version of $\hat{T}_*^{(-k)}$, and the scaling factors are the upper Toeplitz diagonals of $M_*^{(k)}$ and $\hat{M}_*^{(k)}$.

The essence of BPC is that these properties and the Bareiss recursion allow us to perform a *sideways upshift* (SU) of the ZB (an SU-*cycle*) as follows.

* Compute $\hat{T}_*^{(2-j)}$, $\hat{T}_*^{(1-j)}$, and the corresponding $\hat{M}$'s using (3.7). (See Algorithm 3.2, steps 1–6. Here, scale factors $\hat{m}_*^{(1-j)}$ and $\hat{m}_*^{(2-j)}$ have been set to equal $m_*^{(j-1)}$ and $m_*^{(j-2)}$, respectively.)

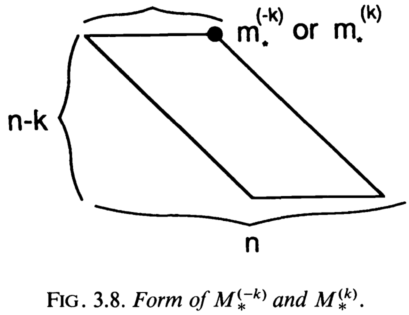


FIG. 3.8. *Form of* $M_*^{(-k)}$ *and* $M_*^{(k)}$.

* Compute $\hat{T}_{*L1}^{(j-2)}$ from $\hat{T}_{*}^{(2-j)}$ and $\hat{T}_{*}^{(1-j)}$ by rearranging the Bareiss step (2.10); compute $\hat{M}_{*L1}^{(j-2)}$ by rearranging the multiplier update step (3.5). (See Algorithm 3.2, steps 7 and 8. The derivation of these steps is detailed in Appendix E.)

* Compute the bottom-left element of $\hat{T}_{*}^{(j-2)}$ using the identity $\hat{M}_{*}^{(j-2)}T = \hat{T}_{*}^{(j-2)}$; the rest of the bottom row of $\hat{T}_{*}^{(j-2)}$ follows from the Toeplitz property (see Algorithm 3.2, step 10).

* Compute $\hat{T}_{*}^{(j-1)}$ and $\hat{M}_{*}^{(j-1)}$ by a normal Bareiss step and a multiplier matrix update step (3.6) (see Algorithm 3.2, steps 11 and 12).

The detailed steps of an SU-cycle are as follows.

ALGORITHM 3.2.

```
1. $\hat{T}_{*}^{(2-j)} := T_{*}^{(j-2)}$,
2. $\hat{M}_{*}^{(2-j)} := M_{*}^{(j-2)}$,
3. $\hat{m}_{*}^{(2-j)} := m_{*}^{(j-2)}$,
4. $\hat{T}_{*}^{(1-j)} := T_{*}^{(j-1)}$,
5. $\hat{M}_{*}^{(1-j)} := M_{*}^{(j-1)}$,
6. $\hat{m}_{*}^{(1-j)} := m_{*}^{(j-1)}$,
7. $\hat{T}_{*L1}^{(j-2)} := \hat{T}_{*F1}^{(2-j)} - (\hat{m}_{*}^{(2-j)}/\hat{m}_{*}^{(1-j)})\hat{T}_{*}^{(1-j)}$,
8. $\hat{M}_{*L1}^{(j-2)} := \hat{M}_{*F1}^{(2-j)} - (\hat{m}_{*}^{(2-j)}/\hat{m}_{*}^{(1-j)})\hat{M}_{*}^{(1-j)}$,
9. $\hat{\mathbf{m}}_{*n-j+2}^{(j-2)} := \hat{\mathbf{m}}_{*n-j+1}^{(j-2)} Z_{-1}^{T}$,
10. $\hat{\mathbf{t}}_{*n-j+2}^{(j-2)} := [\hat{\mathbf{m}}_{*n-j+2}^{(j-2)}\mathbf{t}_{\cdot 1}, \hat{\mathbf{t}}_{*n-j+1,1:n-1}^{(j-2)}]$,
11. $\hat{T}_{*}^{(j-1)} := \hat{T}_{*L1}^{(j-2)} - \hat{m}_{j-1}\hat{T}_{*}^{(1-j)}$, where $\hat{m}_{j-1} = \hat{t}_{A}^{(j-2)}/\hat{t}_{A}^{(1-j)}$,
12. $\hat{M}_{*}^{(j-1)} := \hat{M}_{*L1}^{(j-2)} - \hat{m}_{j-1}\hat{M}_{*}^{(1-j)}$,
```

where $\mathbf{a}_{*i}$ denotes row i of A_* for any A.

The main operations in an SU-cycle are illustrated in Fig. 3.9.

The above procedure requires $4n - j$ operations for each ZB upshift, compared to $4(n-j)$ for a Bareiss step. The extra $3j$ operations are required for steps 8, 10, and 12 of Algorithm 3.2. The procedure for a ZB downshift is similar.

3.7. Elimination of small diagonals. Suppose that p diagonals above the ZB and q diagonals below were rejected as pivots because they were "small" by some criterion (to be discussed later). Suppose that the desired pivot was $t_{p+1}^{(-j)}$, in which case the ZB was moved up by p places. Then $p + q$ diagonals below the ZB will generally be small. To see this, observe that (3.2) can be written using normal matrix notation as follows:

$$(3.8) \qquad \frac{\tilde{t}_{*im}^{(-j)}}{\tilde{t}_{*i'm'}^{(-j)}} = \frac{t_{*im}^{(-j)} + \varepsilon_{im}}{t_{*i'm'}^{(-j)} + \varepsilon_{i'm'}}, \quad 1 \leqq i \leqq n-j, \quad 1 \leqq m \leqq n,$$

where $t_{*i'm'}^{(-j)} = t_{\max}^{(-j)}$, and ε_{im} and $\varepsilon_{i'm'}$ are bounded as in (3.3). Equation (3.8) shows (i) the q diagonals below the original ZB remain small, and (ii) the p new diagonals above these (originally zero) are also small because their values relative to $\tilde{t}_{*i'm'}^{(-j)}$ will be $\varepsilon_{im}/(t_{*i'm'}^{(-j)} + \varepsilon_{i'm'})$.

These $p + q$ small diagonals can be eliminated by using $p + q$ normal Bareiss cycles, henceforth called B-cycles.

Suppose now that the desired pivot was $t_{-q-1}^{(-j)}$, so that the ZB was moved *down* $q + 1$ places, in which case the desired pivot appeared above the ZB (cf. Fig. 3.3). In

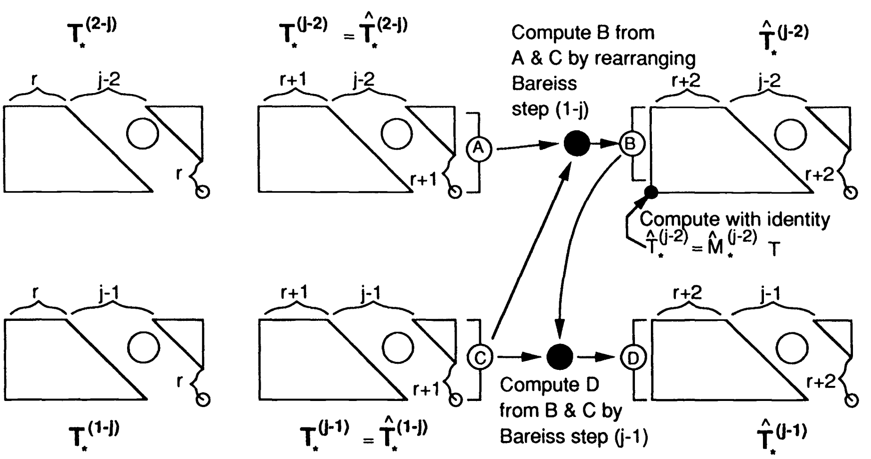


FIG. 3.9. *Main steps in computing* $\hat{T}_*^{\pm(j-2)}$ *and* $\hat{T}_*^{\pm(j-1)}$ *from* $T_*^{\pm(j-2)}$ *and* $T_*^{\pm(j-1)}$ *in a sideways upshift. Corresponding multiplier matrices are computed similarly.*

this case, $p + q$ small diagonals appear above the ZB in $T^{(j)}$. To see this, recall property C above: $T^{(j)}$ for ZBD $= -q - 1$ is a scaled version of $T^{(-j)}$ with ZBD $= -q$. The $p + q$ diagonals above the ZB of this $T^{(-j)}$ are in the same position as either zero or small diagonals of the original $T^{(-j)}$ (with ZBD $= 0$), so by (3.8) they are small compared to the largest diagonal, hence the $p + q$ diagonals above the ZB of $T^{(j)}$ (with ZBD $= q + 1$) are small. These can also be eliminated using $p + q$ B-cycles.

3.8. Restoration of Bareiss form. After a backtrack procedure has been performed and small diagonals eliminated, the ZB should be moved back to its normal position. Recall Fig. 3.2, which shows the normal Bareiss form of $T^{(-j)}$ and the form of $T^{(-j)}$ with ZBD $= p$. To restore the Bareiss form, we carry out p U-cycles if the ZB was displaced down p places, or p D-cycles if the ZB was displaced up p places.

3.9. Nontermination of restoration of Bareiss form: Augmentation of *T*. Occasionally [13] when the pivoting is carried out very late in the reduction, and the ZB displacement, say d, is relatively large in modulus, restoration of Bareiss form (RBF) is not complete at cycle $n - 1$, and the final ZBD, say d_f, is not zero. This occurs if $k + |d| > n - 1$, where k is the cycle number before RBF begins; then $|d_f| = k + |d| - (n - 1)$. Figure 3.10 shows the final reduced rows for ZBDs of 0 and 3, respectively. To complete the reduction, a non-Toeplitz matrix of order $|d_f|$ must be triangularized. In most cases, $|d_f|$ will be small compared to n, and the cost of this reduction will not be significant.

In the occasional case where $|d_f|/n$ is not small, a different technique must be used: if T is *augmented* by $r = k + |d| - (n - 1)$ Toeplitz rows and columns, then RBF can finish, giving the solution of an $(n + r) \times (n + r)$ system. The right-hand side must also

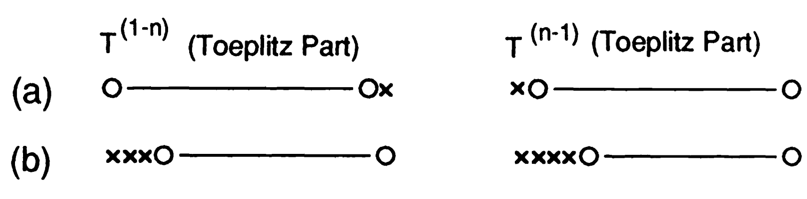


FIG. 3.10. *Toeplitz part of* $T^{(1-n)}$ *and* $T^{(n-1)}$ (a) *with* $d_f = 0$, (b) *with* $d_f = 3$.

be augmented by an r-vector $\mathbf{c}$. It can be seen that if $\mathbf{c}$ is chosen so that $\mathbf{x}_2 = \mathbf{0}$ in the system

$$T_{n+r}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix} = \begin{bmatrix}\mathbf{b}\\ \mathbf{c}\end{bmatrix}, \tag{3.9}$$

where T_{n+r} is the augmented matrix, then $\mathbf{x}_1 = \mathbf{x}$. The problem is to compute $\mathbf{c}$ in $O(n^2)$ operations. Here, one method is used when RBF is carried out using SU-cycles or sideways downshift (SD) cycles, and another when RBF is performed with U- or D-cycles. In the first case, it can be shown that the last r cycles of the reduction of T_{n+r} are B-cycles, and that

$$b_{n+i}^{(1-n-r)} = b_{n+i}^{(1-n-i)} = \mathbf{m}_{*1,1:n+i}^{(1-n-i)}[\mathbf{b}^T, \mathbf{c}_{1:i}^T]^T, \qquad i = 1, \ldots, r.$$

For any given $\mathbf{c}$, it can be shown that $\mathbf{x}_2$ can be computed by solving the upper-triangular system

$$\mathbf{t}_{*1,n+1:n+r}^{(1-n-i)}\mathbf{x}_2 = b_{n+i}^{(1-n-i)}, \qquad i = 1, \ldots, r,$$

so if $\mathbf{c}$ is selected such that $b_{n+i}^{(1-n-i)} = 0$, $i = 1, \ldots, r$, then $\mathbf{x}_2 = 0$. This can be ensured by selecting each c_i such that

$$\mathbf{m}_{*1,1:n+i}^{(1-n-i)}[\mathbf{b}^T, \mathbf{c}_{1:i}^T]^T = 0, \qquad i = 1, \ldots, r,$$

which requires $O(nr)$ operations.

In the case where RBF is performed using U- or D-cycles, the computation of $\mathbf{c}$ in $O(nr)$ operations is more involved. Details are given in [13].

3.10. Extraction of upper triangle for the solution of (2.1). By considering the zero-patterns in the Toeplitz parts of the $T^{(-j)}$ (e.g., see Figs. 3.2 and 3.3), a column-permuted upper-triangular system can be extracted by selecting appropriate rows from the $T^{(-j)}$ and $\mathbf{b}^{(-j)}$. Observe that if SU- or SD-cycles are carried out at iteration j, there will be several values of $T^{(-j)}$, one for each ZBD. We therefore denote $T^{(-j,d)}$ as the value of $T^{(-j)}$ at ZBD d in the following. We will show that the following algorithm produces a column-permuted upper triangle U'.

ALGORITHM 3.3.

```
1. $\mathbf{u}'_{n\cdot} := \mathbf{t}_{*1\cdot}^{(1-n,0)}$; $h := \bar{h} := 0$; $i := n - 1$ {$i$: index of *next* row of $U'$};
2. **For** each iteration from the last *back* to the first, **do**
2.1. **case** cycle-type **of**
         $U : h := h - 1$,
         $D,SD : h := h + 1$;
2.2. $\bar{h} := \max(\bar{h}, h)$;
2.3. $a := 1 + \bar{h} - h$;
2.4. **if** $\mathbf{t}_{*a\cdot}^{(-j,d)}$ has nonzero entry $t_{*am}^{(-j,d)}$ for which $u'_{lm} = 0, i < l \leqq n$ **then**
         $\mathbf{u}'_{i\cdot} := \mathbf{t}_{*a\cdot}^{(-j,d)}$,
         $i := i - 1$.
```

The following lemma shows that (i) $\mathbf{t}_{*a\cdot}^{(-j,d)}$ always exists and (ii) U' is a column-permuted upper triangle.

LEMMA 3.3. *For each cycle of the* PBA, *let a and $\mathbf{u}'_{i\cdot}$ be as in Algorithm* 3.3. *Then*

A. $1 \leqq a \leqq n - j$ *for each iteration j and* ZBD d.

B. $\mathbf{u}'_{i\cdot}$ *has at exactly one nonzero entry u'_{im} for which $u'_{lm} = 0$, $i < l \leqq n$.*

C. *The last computed row of U' is $\mathbf{u}'_{1\cdot}$, and U' is an $n \times n$ column-permuted upper triangle.*

Proof. The first inequality in property A follows immediately from step 2.2. The second follows from the fact that there cannot be more than $n-j-1$ U-cycles after cycle j, and hence $\bar{h}$ cannot exceed h by more than $n-j-1$. For property B, it can be easily verified from the zero-patterns in the $T_*^{(-j,d)}$ that $\mathbf{t}_{*a\cdot}^{(-j,d)}$ has *at most* one nonzero entry $t_{*am}^{(-j,d)}$ for which $u'_{lm}=0$, $i<l\leqq n$, hence, by step 2.4, $\mathbf{u}'_{i\cdot}$ has *exactly* one nonzero entry u'_{im} for which $u'_{lm}=0$, $i<l\leqq n$. For property C, observe that the last computed row $\mathbf{u}'_{i\cdot}$ is $t_{*a\cdot}^{(-0,d)}$. By applying property B successively to $\mathbf{u}'_{n-1\cdot}, \dots, \mathbf{u}'_{i\cdot}$ and noting that $t_{*a\cdot}^{(-0,d)}$ has no zeros, we obtain property C. □

Remark. The test in step 2.4 can be carried out efficiently by maintaining an array of flags, the mth entry of which is set when a new nonzero entry $t_{*am}^{(-j,d)}$ is found. The updated array is then used in the test in the next iteration of the loop.

3.11. Summary of the composite pivoting procedure. We execute the Bareiss algorithm (B-cycles) until $t_A^{(-j)}$ is less than some threshold (discussed below), move the ZB by a backtrack procedure until an acceptable pivot is found, eliminate any small diagonals using B-cycles, then restore to Bareiss form. This procedure is repeated when necessary. Processing continues to cycle $n-1$. The permuted upper triangle is extracted using Algorithm 3.3.

3.12. Pivot selection strategy. We have not yet given a criterion for selecting a pivot that is "large enough." The following heuristic strategy has been used to handle cases where groups of badly conditioned leading submatrices are bracketed by well-conditioned leading submatrices.

Case 1. Use of BPA or BPB to handle contiguous NSLSs. We execute the Bareiss algorithm until $|t_A^{(-j)}|/|t_{\max}^{(-j)}|<0.001$, and then provisionally select the nearest pivot $t_P^{(-j)}$ to the ZB satisfying $|t_P^{(-j)}|/|t_{\max}^{(-j)}|\geqq 0.001$. However, by Lemma 3.2 the execution of BPA or BPB to move the ZB could change $t_P^{(-j)}$ significantly unless $|t_P^{(-j)}|$ is considerably larger in modulus than elements between it and the ZB. Hence we also impose the condition that $|t_P^{(-j)}|>K\|\mathbf{t}_{ZP}^{(-j)}\|_1$, where K is a moderately large number, say 10, and $\mathbf{t}_{ZP}^{(-j)}$ is the vector containing the entries between the ZB and $t_P^{(-j)}$. If there is no $t_P^{(-j)}$ that satisfies the latter condition, we select $t_P^{(-j)}$ such that $|t_P^{(-j)}|/\|\mathbf{t}_{ZP}^{(-j)}\|_1$ is maximized.

Case 2. Use of BPC to handle contiguous NSLSs. As for Case 1, we pivot if $|t_A^{(-j)}|/|t_{\max}^{(-j)}|<0.001$. We then move the ZB up or down until $|t_A^{(-j)}|$ or $|t_B^{(-j)}|$, respectively, is greater than $0.001\,|t_{\max}^{(-j)}|$. This strategy is simpler than for Case 1 because the pivots can be monitored at each cycle in BPC. However, BPC is slower than BPA or BPB.

General remarks. A higher threshold, say $|t_A^{(-j)}|/|t_{\max}^{(-j)}|<0.1$, could be used to handle cases of mildly ill conditioned leading submatrices. This should improve the overall performance of the Bareiss algorithm, though at the expense of extra backtracking and pivoting. Further work is required on testing the algorithm using various thresholds on a large variety of Toeplitz matrices to determine the improvement achievable and to measure typical pivoting overheads.

3.13. Remark on error propagation in the pivoted Bareiss algorithm. The error analysis of § 2.4 was generalized to the pivoted case in [13], and the result is analogous to that of (2.21). Hence, if pivots are to be selected to be as large as possible, the multipliers will be kept small, minimizing the rounding error propagation.

3.14. Numerical results. The Bareiss algorithm was run on the system $T\mathbf{x}=\mathbf{b}$ with

$$\mathbf{t}_{1\cdot}=[5,-1,6,2,5.697,5.850,3,-5,-2,-7,1,10,-15], \tag{3.10a}$$

$$\mathbf{t}_{\cdot 1}=[5,1,-3,12.755,-19.656,28.361,-7,-1,2,1,-6,1,-0.5]^T, \tag{3.10b}$$

$$\mathbf{b}=T[1,\dots,1]^T. \tag{3.10c}$$

In this matrix, a change of only 1.5×10^{-3} in four elements is required to make T_4, T_5, T_6, and T_7 singular. The maximum relative error in $\mathbf{x}$ using 56 bits of precision was 13 percent. When the pivoted algorithm was run on the same matrix, the maximum relative error was 7×10^{-14}.

4. The pivoted Bareiss factorizer. It can easily be shown for the Bareiss algorithm that the multiplier matrices $M^{(\pm j)}$ (3.4) are triangular with bandwidth $j + 1$, and hence, using the persymmetry of T, that the upper and lower triangular factors of T are given, respectively by,

$$U = T^{(1-n)}, \tag{4.1}$$

$$L = T^{(n-1)T2}/t_{11}. \tag{4.2}$$

An upper- and lower-triangular matrix, whose product is a permuted version of T, may be extracted from a restricted pivoted Bareiss algorithm which contains only B-, U-, and D-cycles. The algorithm is as follows:

```
ALGORITHM 4.1.
  1. Form the permutation vectors s and v as follows:
  1.1.
(4.3a, b) $s_1 := n_U + 1$;  $v_1 := n_D + 1$  {$n_U, n_D$ : no of U- and D-cycles respectively},
  1.2. for $j := 1$ to $n - 1$ do
         if cycle $j$ is a U-cycle then
  1.2A.
(4.4a, b)      $s_{j+1} := \min\{s_1, \ldots, s_j\} - 1$,      $v_{j+1} := \max\{v_1, \ldots, v_j\} + 1$,
         elseif cycle $j$ is a B-cycle then
  1.2B.
(4.5a, b)      $s_{j+1} := \max\{s_1, \ldots, s_j\} + 1$,      $v_{j+1} := \max\{v_1, \ldots, v_j\} + 1$,
         else {cycle $j$ is a D-cycle}
  1.2C.
(4.6a, b)      $s_{j+1} := \max\{s_1, \ldots, s_j\} + 1$,      $v_{j+1} := \min\{v_1, \ldots, v_j\} - 1$;
  2. Form the permutation matrices
(4.7a, b) $P : \{p_{ij} = \delta_{s_i,j}\}$  and  $Q : \{q_{ij} = \delta_{v_i,j}\}$;
  3. Set
(4.8a, b) $g := n_U$,      $h := n_D$;
  4.
(4.9a–d)  $\mathbf{l}'_{\cdot 1} := P\mathbf{t}^{RT}_{n-n_D\cdot}$,   $\mathbf{l}_{\cdot 1} := \mathbf{l}'_{\cdot 1}/l'_{11}$,   $\mathbf{u}_{1\cdot} := \mathbf{t}_{1+n_U\cdot}Q^T$,   $m^{(\pm 0)}_{UT} := m^{(\pm 0)}_{LT} := 1$,
     where $m^{(\pm j)}_{UT}$ ($\equiv m^{(\pm j)}_{*}$, see § 3.6) and $m^{(\pm j)}_{LT}$ are the value of the upper and lower
  Toeplitz diagonals, respectively, of $M^{(\pm j)}$.
  5. for $j := 1$ to $n - 1$ do
       if cycle $j$ is a U-cycle then
  5A.
(4.10a, b)   $m^{(-j)}_{UT} := m^{(1-j)}_{UT}$,      $m^{(-j)}_{LT} := -m_{-j}m^{(j-1)}_{LT}$,
(4.10c, d)   $m^{(j)}_{UT} := m^{(j-1)}_{UT} - m_j m^{(-j)}_{UT}$,      $m^{(j)}_{LT} := -m_j m^{(-j)}_{LT}$,
```

(4.10e, f) $\mathbf{l}'_{\cdot j+1} := P\mathbf{t}^{(-j)RT}_{b(g)\cdot}, \qquad \mathbf{l}_{\cdot j+1} := \mathbf{l}'_{\cdot j+1}/l'_{j+1,j+1},$

(4.10g) $\mathbf{u}_{j+1\cdot} := \mathbf{t}^{(-j)}_{a(h)\cdot}Q^T/m^{(-j)}_{LT},$ decrement g

{here $a(i)$ and $b(i)$ are the rows with i nonzero elements to the left and right of the ZB, respectively},

elseif cycle j is a B-cycle **then**

5B.

(4.11a, b) $m^{(-j)}_{UT} := m^{(1-j)}_{UT}, \qquad m^{(-j)}_{LT} := -m_{-j}m^{(j-1)}_{LT},$

(4.11c, d) $m^{(j)}_{UT} := -m_j m^{(-j)}_{UT}, \qquad m^{(j)}_{LT} := m^{(j-1)}_{LT} - m_j m^{(-j)}_{LT},$

(4.11e–g) $\mathbf{l}'_{\cdot j+1} := P\mathbf{t}^{(j)RT}_{b(g)\cdot}, \quad \mathbf{l}_{\cdot j+1} := \mathbf{l}'_{\cdot j+1}/l'_{j+1,j+1}, \quad \mathbf{u}_{j+1\cdot} := \mathbf{t}^{(-j)}_{a(h)\cdot}Q^T/m^{(-j)}_{UT},$

else {cycle j is a D-cycle};

5C.

(4.12a, b) $m^{(j)}_{UT} := -m_j m^{(1-j)}_{UT}, \qquad m^{(j)}_{LT} := m^{(j-1)}_{LT},$

(4.12c, d) $m^{(-j)}_{UT} := -m_{-j}m^{(j)}_{UT}, \qquad m^{(-j)}_{LT} := m^{(1-j)}_{LT} - m_{-j}m^{(j)}_{LT},$

(4.12e, f) $\mathbf{l}'_{\cdot j+1} := P\mathbf{t}^{(j)RT}_{b(g)\cdot}, \qquad \mathbf{l}_{\cdot j+1} := \mathbf{l}'_{\cdot j+1}/l'_{j+1,j+1},$

(4.12g) $\mathbf{u}_{j+1\cdot} := \mathbf{t}^{(j)}_{a(h)\cdot}Q^T/m^{(j)}_{UT},$ decrement h.

We then have the following theorem.

THEOREM 4.1. *Let $T^{(\pm j)}$ be computed by a pivoting sequence consisting only of* B-, U-, *and* D-*cycles*; *let P and Q be the permutation matrices*; *and let L and U be the matrices as computed from $T^{(\pm j)}$ by Algorithm* 4.1. *Then* (i) *L is unit lower triangular and U is upper triangular, and* (ii) $LU = PTQ^T$.

Outline of proof. The recursions for $m^{(\pm j)}_{UT}$ and $m^{(\pm j)}_{LT}$ are easily verified using the multiplier matrix recursions (3.5) and (3.6). By considering the zero-patterns in the Toeplitz parts of $T^{(-j)}$ and $T^{(j)}$ from which the $\mathbf{t}^{(\pm j)}_{a(h)\cdot}$ and $\mathbf{t}^{(\pm j)RT}_{b(g)\cdot}$ are selected, it can be verified that these rows and columns, after being permuted by Q and P^T, respectively, form upper and lower triangles. It can also be shown for B- and U-cycles that (i) $\mathbf{t}^{(-j)}_{a(h)\cdot}$ is a linear combination of $\{\mathbf{t}_{i\cdot}\}^{s_{j+1}}_{s_1}$, and (ii) after normalization by $m^{(-j)}_{UT}$ (B-cycles) or $m^{(-j)}_{LT}$ (U-cycles), $\mathbf{u}_{j+1\cdot}Q$ is a *normalized* linear combination (NLC) of $\{\mathbf{t}_{i\cdot}\}^{s_{j+1}}_{s_1}$ (coefficient of $\mathbf{t}_{s_{j+1}\cdot}$ is unity). Hence, from the definition of P, $\mathbf{u}_{j+1\cdot}Q$ is a NLC of $\{(PT)_{i\cdot}\}^{j+1}_1$, because premultiplication by P maps rows $s_1, s_2, \ldots,$ to rows 1, 2, ... Thus $\mathbf{u}_{j+1\cdot}$ is a NLC of $\{(PTQ^T)_{i\cdot}\}^{j+1}_1$, i.e., $\mathbf{u}_{j+1\cdot}$ is row $j+1$ of the unique U-factor of PTQ^T. The proof for D-cycles is similar. We can similarly show that $\mathbf{l}_{\cdot j+1}$ is column $j+1$ of the unique L-factor of PTQ^T. Full details are given in Appendix F.

Numerical results. The system $T\mathbf{x} = \mathbf{b}$ specified by (3.11) was factored using the Bareiss algorithm and (4.1) and (4.2). The largest element in the error matrix $\bar{L}\bar{U} - T$ had a modulus of 2.14. When T was factored using PBA together with Algorithm 4.1, the largest element in the error matrix $P\bar{L}\bar{U}Q^T - T$ had modulus 9×10^{-13}.

5. The pivoted Trench algorithm. We first show how the Trench algorithm may be derived from the multiplier matrices produced from a variant of the Bareiss algorithm, and in the same way, indicate how to derive a pivoted Trench algorithm from PBA.

5.1. Derivation of the Trench algorithm from multiplier matrices. Bareiss also proposed a "symmetric" version of his algorithm, viz.,

for $j := 1$ **to** $n-1$ **do**

(5.1) $T^{(-j)} := T^{(1-j)} - m_{-j}Z_{-j}T^{(j-1)},$ where $m_{-j} = t^{(1-j)}_{j+1,1}/t^{(j-1)}_{11},$

(5.2) $T^{(j)} := T^{(j-1)} - m_j Z_j T^{(1-j)},$ where $m_j = t^{(j-1)}_{n-j,n}/t^{(1-j)}_{nn}.$

The recursions for $\mathbf{b}^{(\pm j)}$ are omitted here. The recursion for $T^{(-j)}$ is the same as in (2.4) and (2.5), but the recursion for $T^{(j)}$, unlike that in (2.6) and (2.7), has the same form as the recursion for $T^{(-j)}$, apart from the signs of the iteration numbers. Hence the term "symmetric." The multiplier matrix recursion is

$$M^{(-j)} := M^{(1-j)} - m_{-j}Z_{-j}M^{(j-1)}, \tag{5.3a}$$

$$M^{(j)} := M^{(j-1)} - m_jZ_jM^{(1-j)}. \tag{5.3b}$$

Now $M^{(-j)}$ is $(j+1)$-diagonal lower triangular, and it is easily seen that rows $j+1$ to n are Toeplitz. Similarly, $M^{(j)}$ is $(j+1)$-diagonal upper triangular, with rows 1 to $n-j$ being Toeplitz. Define

$$\mathbf{m}_T^{(-j)} \equiv [m_{j+1,1}^{(-j)}, \ldots, m_{j+1,j+1}^{(-j)}] \quad \text{and} \quad \mathbf{m}_T^{(j)} \equiv [m_{11}^{(j)}, \ldots, m_{1,j+1}^{(j)}].$$

Note that the right-hand sides are the first rows of the Toeplitz parts of $M^{(-j)}$ and $M^{(j)}$, respectively. Then from (5.3) we have

$$\mathbf{m}_T^{(-j)} = [0, \mathbf{m}_T^{(1-j)}] - m_{-j}[\mathbf{m}_T^{(j-1)}, 0]; \ \mathbf{m}_T^{(j)} = [\mathbf{m}_T^{(j-1)}, 0] - m_j[0, \mathbf{m}_T^{(j-1)}]. \tag{5.4}$$

The recursion (5.4) would be self-contained if we have expressions or recursions for the numerators and denominators of the $m_{\pm j}$ in (5.1) and (5.2). By the symmetric Bareiss recursion and the definition of the multiplier matrices, we have

$$t_{j+1,1}^{(1-j)} = \mathbf{m}_T^{(1-j)}[t_{21}, \ldots, t_{j+1,1}]^T, \tag{5.5a}$$

$$t_{11}^{(j-1)} = t_{11}^{(j-2)} - m_{j-1}t_{j1}^{(2-j)}, \tag{5.5b}$$

$$m_{-j} = t_{j+1,1}^{(1-j)}/t_{11}^{(j-1)} = t_B^{(1-j)}/t_B^{(j-1)}, \tag{5.5c, d}$$

$$t_{n-j,n}^{(j-1)} = \mathbf{m}_T^{(j-1)}[t_{n-j,n}, \ldots, t_{n-1,n}]^T, \tag{5.6a}$$

$$t_{nn}^{(1-j)} = t_{nn}^{(2-j)} - m_{1-j}t_{n-j+1,n}^{(j-2)}, \tag{5.6b}$$

$$m_j = t_{n-j,n}^{(j-1)}/t_{nn}^{(1-j)} = t_A^{(j-1)}/t_A^{(1-j)}. \tag{5.6c, d}$$

The recursion (5.4)–(5.6) yields $\mathbf{m}_T^{(1-n)}$ and $\mathbf{m}_T^{(n-1)}$ in $n-1$ cycles. However,

$$\mathbf{m}_T^{(1-n)}T = [0, \ldots, 0, t_{nn}^{(1-n)}];$$

therefore,

$$\mathbf{m}_T^{(1-n)}/t_{nn}^{(1-n)} = (T^{-1})_{n\cdot} = (T^{-1})_{\cdot 1}^{RT} \quad \text{by persymmetry,} \tag{5.7}$$

and similarly,

$$\mathbf{m}_T^{(n-1)}/t_{11}^{(n-1)} = (T^{-1})_{1\cdot}. \tag{5.8}$$

Thus the recursion (5.4)–(5.6) produces the first row and column of T^{-1}; by a suitable change in notation, this recursion can be shown to be equivalent to Phase I of the Trench algorithm [14].

5.2. Derivation of pivoted Trench algorithm from PBA. A pivoted Trench algorithm may be derived from any PBA variant in the same way that the Trench algorithm was derived from the multiplier matrices of the Bareiss algorithm.

The general methodology is as follows.

1. Write the multiplier matrix recursion, analogous to (5.3). Each cycle of this recursion is the same as in PBA, with $M^{(\pm j)}$ replacing $T^{(\pm j)}$. Recursions for the various types of pivoting cycles are given in (2.9)–(2.12) (B-cycles) and §§ 3.2, 3.5, and 3.6 (U-, D-, and SU- and SD-cycles). They all have the form

$$M_*^{(\pm j)} := \tilde{Z}_u M_*^{\pm(j-1)} - m_{\pm j}\tilde{Z}_v M_*^{\mp(j-s)}, \tag{5.9}$$

TABLE 5.1
Numerators and denominators of multipliers for various pivoting operations.

	Bareiss symmetric algorithm	Algorithm 3.1	U-cycles	B-cycles	D-cycles
n_{-j}	$t_B^{(1-j)}$	$t_A^{(1-j)}$	$t_A^{(1-j)}$	$t_B^{(1-j)}$	$t_B^{(1-j)}$
d_{-j}	$t_B^{(j-1)}$	$t_A^{(j-1)}$	$t_A^{(j-1)}$	$t_B^{(j-1)}$	$t_B^{(j)}$
n_j	$t_A^{(j-1)}$	$t_B^{(j-1)}$	$t_A^{(j-1)}$	$t_A^{(j-1)}$	$t_B^{(j-1)}$
d_j	$t_A^{(1-j)}$	$t_B^{(-j)}$	$t_A^{(-j)}$	$t_A^{(-j)}$	$t_B^{(1-j)}$

where $s = 0$ or 1, and for any k, $\tilde{Z}_k$ is a rectangular matrix that selects all but the first k rows (for k positive) or the last k rows (for k negative) of the matrix it premultiplies. u and v depend on the cycle type and the sign of $\pm j$. For example, in a U-cycle, $u = 1$ and $v = -1$ for steps $(-j)$, and $u = 1$ and $v = 0$ for steps $(+j)$. The initialization step of the recursion is $M^{(-0)} := M^{(+0)} := I$.

2. Write the recursion corresponding to (5.4) for $\mathbf{m}_T^{(\pm j)}$, the nonzero elements of the first rows of the Toeplitz parts of $M^{(\pm j)}$. The recursions for the $\mathbf{m}_T^{(\pm j)}$ can be shown from (5.9) to have the form

$$\mathbf{m}_T^{(\pm j)} := [\mathbf{0}_{u^+}, \mathbf{m}_T^{\pm(j-1)}, \mathbf{0}_{u^-}] - m_{\pm j}[\mathbf{0}_{v^+}, \mathbf{m}_T^{\mp(j-s)}, \mathbf{0}_{v^-}],$$

where u, v, and s are as in (5.9), $u^+ = \max(0, u)$, $u^- = \max(0, -u)$, $v^\pm$ are defined similarly, and $\mathbf{0}_{u^\pm}$ and $\mathbf{0}_{v^\pm}$ are zero vectors of length $u^\pm$ and $v^\pm$, respectively.

3. Write, in analogy to (5.5d) and (5.6d), expressions for $m_{\pm j}$ in the form

$$m_{\pm j} = n_{\pm j}/d_{\pm j},$$

where $n_{\pm j}$, the numerator term, and $d_{\pm j}$, the denominator term, are one of $t_A^{\pm(j-s)}$ or $t_B^{\pm(j-s)}$, $s = 0$ or 1. Table 5.1 shows the $n_{\pm j}$ and $d_{\pm j}$ for the various types of pivoting steps, as detailed in § 3.

Depending on the type of pivoting step, these may be computed either from previous iterates using the Bareiss recursion, or by application of the identities

$$\mathbf{m}_{k\cdot}^{\pm(j-s)}\mathbf{t}_{\cdot l} = \mathbf{t}_{kl}^{\pm(j-s)}. \tag{5.10}$$

In the latter case, $\mathbf{m}_{k\cdot}^{\pm(j-s)}$ is a row of $M_*^{\pm(j-s)}$ and has the form $[\mathbf{0}_p, \mathbf{m}_T^{\pm(j-s)}, \mathbf{0}_q]$ where p and q are positive integers (see Fig. 3.8). Details for the different types of pivoting are given in Table 5.2.

TABLE 5.2
Calculation of $t_A^{(\pm j)}$ and $t_B^{(\pm j)}$ for various pivoting operations.

	Bareiss symmetric algorithm	Algorithm 3.1	U-cycles	B-cycles	D-cycles
$t_A^{(-j)}$	$t_A^{(1-j)} - m_{-j}t_A^{(j-1)}$	use (5.10)	use (5.10)	$t_A^{(1-j)} - m_{-j}t_A^{(j-1)}$	$t_A^{(1-j)}$
$t_B^{(-j)}$	use (5.10)	$t_B^{(1-j)} - m_{-j}t_B^{(j-1)}$	$t_B^{(1-j)} - m_{-j}t_B^{(j-1)}$	use (5.10)	use (5.10)
$t_A^{(j)}$	use (5.10)	$t_A^{(j-1)}$	use (5.10)	use (5.10)	$t_A^{(j-1)} - m_jt_A^{(1-j)}$
$t_B^{(j)}$	$t_B^{(j-1)} - m_jt_B^{(1-j)}$	use (5.10)	$t_B^{(j-1)}$	$t_B^{(j-1)}$	use (5.10)

4. The first row and column of T^{-1} are given by (5.7) and (5.8). The remainder of T^{-1} is computed using Phase II of the Trench algorithm.

5.3. Numerical results. The Trench algorithm was run on a Toeplitz matrix with first row and column

$$\mathbf{t}_{1\cdot} = (8, 4, 1, 6, 2, 3), \qquad \mathbf{t}_{\cdot 1} = (8, 4, -34 + 5 \times 10^{-13}, 5, 3, 1)^T.$$

T would be singular if $t_{31} = -34$. For the Trench algorithm, $\|\overline{TT^{-1}} - I\|$ was 0.52. A pivoted Trench algorithm derived from Algorithm 3.1 was then run on T, and the value of $\|\overline{TT^{-1}} - I\|$ was 6.4×10^{-15} in this case.

6. Conclusions. The Bareiss algorithm and related $O(n^2)$ algorithms can suffer a large loss in accuracy if some leading submatrices are ill conditioned. Some pivoting schemes have been proposed for the Bareiss algorithm to handle these cases. These schemes can also handle cases of exact singularity, because exactly singular leading submatrices cause extra zero diagonals in $T^{(-j)}$, rather than small diagonals. Numerical singularities (exact singularities perturbed by rounding errors) cause very small diagonals in $T^{(-j)}$. The heuristic pivoting strategy used would be expected to handle cases where groups of ill-conditioned leading submatrices are bracketed by well-conditioned submatrices. This case is similar to a Padé approximation [1] where "blocks" of identical entries are encountered in computing a row of a nonnormal Padé table. These blocks correspond to groups of singular contiguous Toeplitz submatrices. Groups of near-singular contiguous Toeplitz submatrices would be expected to yield "pseudo-blocks" of very similar, rather than identical, entries in the Padé table. In either the singular or the near-singular case, the proposed algorithm yields paths which detour around the blocks.

The present algorithm would be expected to perform better than the original Bareiss algorithm in most cases. It is an open question whether there is a pivoting strategy that *optimizes* numerical performance by selecting a suitable path through the Padé table, and that does not incur excessive pivoting overheads.

Appendix A. Proof of (2.14b). The proof is as in [11, pp. 58–59]. Partition T_{j+1} and T_{j+1}^{-1} as follows:

$$T_{j+1} = \begin{bmatrix} T_j & \mathbf{v} \\ \mathbf{u}^T & t_{11} \end{bmatrix}, \qquad T_{j+1}^{-1} = \begin{bmatrix} E & \mathbf{f} \\ \mathbf{g}^T & h \end{bmatrix}.$$

From the identity $T_{j+1}T_{j+1}^{-1} = I$, we have $\mathbf{f}/h = -T_j^{-1}\mathbf{v}$, so

$$\text{(A.1)} \qquad \|\mathbf{f}/h\| \leqq \|T_j^{-1}\|\|\mathbf{v}\| \leqq \|T_j^{-1}\|\|T_{j+1}\| = \beta\kappa(T_j),$$

where

$$\beta \equiv \|T_{j+1}\|/\|T_j\|.$$

Similarly,

$$\text{(A.2)} \qquad \|\mathbf{g}/h\| \leqq \beta\kappa(T_j).$$

The Schur complement of h is $T_j^{-1} = E - \mathbf{fg}^T/h$, whence

$$T_{j+1}^{-1} = \begin{bmatrix} T_j^{-1} & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix} + h\begin{bmatrix} \mathbf{f}/h \\ 1 \end{bmatrix}[\mathbf{g}^T/h, 1].$$

Therefore, using (A.1) and (A.2) we get

$$\text{(A.3)} \qquad \|T_{j+1}^{-1}\| \leqq \|T_j^{-1}\| + h(1 + \beta\kappa(T_j))^2.$$

It is easily shown that after step k of the Bareiss algorithm,

$$h = (t_{nn}^{(-j)})^{-1} = (t_A^{(-j)})^{-1}, \quad \text{using (2.8a)}. \tag{A.4}$$

In the case $\|T_{j+1}^{-1}\| > \|T_j^{-1}\|$, we get from (A.3) and (A.4)

$$|t_A^{(-j)}| \leqq \frac{(1+\beta\kappa(T_j))^2}{\|T_{j+1}^{-1}\| - \|T_j^{-1}\|}$$

and the result follows.

Appendix B. Backward error analysis of the Bareiss algorithm: Proof of (2.17)–(2.21). The development follows that of [11, pp. 71–78]. Define $\hat{T}^{(-j)}$, $\hat{T}^{(j)}$, $\hat{\mathbf{b}}^{(-j)}$, and $\hat{\mathbf{b}}^{(j)}$ to be the result of applying the Bareiss recursions (2.3)–(2.7) without rounding error, but with the $m_{\pm i}$ replaced by their *computed* values $\bar{m}_{\pm i}$. Assume that T is nonsingular, and define the *multiplier matrices* $M^{(\pm i)}$ and $\hat{M}^{(\pm i)}$ by

$$\begin{aligned} &(\text{B.1a}) \qquad M^{(\pm i)}T \equiv T^{(\pm i)} \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{for } i = 0, \ldots . \\ &(\text{B.1b}) \qquad \hat{M}^{(\pm i)}T \equiv \hat{T}^{(\pm i)} \end{aligned}$$

Putting (B.1a) into the Bareiss recursion, we get

$$M^{(-j)} := M^{(1-j)} - m_{-j}Z_{-j}M^{(j-1)}, \tag{B.2}$$

$$M^{(j)} := M^{(j-1)} - m_jZ_jM^{(-j)}. \tag{B.3}$$

It is easily seen that $M^{(-i)}$ is unit lower triangular and $M^{(i)}$ is upper triangular. Clearly $\hat{M}^{(-i)}$ and $\hat{M}^{(i)}$ are unit lower triangular and upper triangular, respectively.

Let $\bar{T}^{(1-n)}$, $\mathbf{x}_u$, $\bar{\mathbf{b}}^{(1-n)}$, $\bar{T}^{(n-1)}$, $\mathbf{x}_l$, and $\bar{\mathbf{b}}^{(n-1)}$ be as in (2.16). Define

$$F \equiv \bar{T}^{(1-n)} - \hat{T}^{(1-n)}, \quad \mathbf{d} \equiv \bar{\mathbf{b}}^{(1-n)} - \hat{\mathbf{b}}^{(1-n)},$$

$$R \equiv \bar{T}^{(n-1)} - \hat{T}^{(n-1)}, \quad \text{and} \quad \mathbf{g} \equiv \bar{\mathbf{b}}^{(n-1)} - \hat{\mathbf{b}}^{(n-1)}.$$

It is clear from the definition of $\hat{\mathbf{b}}^{(\pm i)}$ that

$$\hat{M}^{(\pm i)}\mathbf{b} = \hat{\mathbf{b}}^{(\pm i)}. \tag{B.4}$$

Using the definitions of F and $\mathbf{d}$ in (2.17), premultiplying by $(\hat{M}^{(1-n)})^{-1}$, and using (B.1b) and (B.4), we get

$$(T + (\hat{M}^{(1-n)})^{-1}F)\bar{\mathbf{x}}_u = \mathbf{b} + (\hat{M}^{(1-n)})^{-1}\mathbf{d}. \tag{B.5}$$

Let $T = LU$, where L is unit lower triangular and U is upper triangular. Clearly $L = (M^{(1-n)})^{-1}$ since $M^{(1-n)}$ is unit lower triangular. Thus (B.5) can be written

$$(T + (L + \delta L)F)\bar{\mathbf{x}}_u = \mathbf{b} + (L + \delta L)\mathbf{d}, \tag{B.6}$$

where

$$\delta L = (\hat{M}^{(1-n)})^{-1} - (M^{(1-n)})^{-1}.$$

In a similar manner, and using the persymmetry of T and the fact that $M^{(n-1)}$ is upper triangular, we can show that

$$(T + (U + \delta U)^{T2}R/t_{11})\bar{\mathbf{x}}_l = \mathbf{b} + (U + \delta U)^{T2}\mathbf{g}/t_{11}, \tag{B.7}$$

where

$$\delta U^{T2} = t_{11}[(\hat{M}^{(n-1)})^{-1} - (M^{(n-1)})^{-1}],$$

and $T2$ denotes transpose about the secondary diagonal. Equations (B.6) and (B.7) are equations (2.17) and (2.18), the first two of the equations to be proved. We now derive bounds on $\|F\|$, $\|\mathbf{d}\|$, $\|R\|$, and $\|\mathbf{g}\|$. First we need the following lemma.

LEMMA B.1.

$$F = \sum_{p=1}^{n-1} (N_p^{(1-n)}E^{(p)} + N_{-p}^{(1-n)}E^{(-p)}), \tag{B.8a}$$

$$\mathbf{d} = \sum_{p=1}^{n-1} (N_p^{(1-n)}\mathbf{h}^{(p)} + N_{-p}^{(1-n)}\mathbf{h}^{(-p)}), \tag{B.8b}$$

$$R = \sum_{p=1}^{n-1} (N_p^{(n-1)}E^{(p)} + N_{-p}^{(n-1)}E^{(-p)}), \tag{B.9a}$$

$$\mathbf{g} = \sum_{p=1}^{n-1} (N_p^{(n-1)}\mathbf{h}^{(p)} + N_{-p}^{(n-1)}\mathbf{h}^{(-p)}), \tag{B.9b}$$

where $E^{(\pm p)}$ and $\mathbf{h}^{(\pm p)}$, the local errors incurred in calculating $\bar{T}^{(\pm p)}$ and $\bar{\mathbf{b}}^{(\pm p)}$, are given by

$$\bar{T}^{(-p)} = \bar{T}^{(1-p)} - \bar{m}_{-p}Z_{-p}\bar{T}^{(p-1)} + E^{(-p)}, \tag{B.10a}$$

$$\bar{T}^{(p)} = \bar{T}^{(p-1)} - \bar{m}_pZ_p\bar{T}^{(-p)} + E^{(p)}, \tag{B.10b}$$

$$\bar{\mathbf{b}}^{(-p)} = \bar{\mathbf{b}}^{(1-p)} - \bar{m}_{-p}Z_{-p}\bar{\mathbf{b}}^{(p-1)} + \mathbf{h}^{(-p)}, \tag{B.11a}$$

$$\bar{\mathbf{b}}^{(p)} = \bar{\mathbf{b}}^{(p-1)} - \bar{m}_pZ_p\bar{\mathbf{b}}^{(-p)} + \mathbf{h}^{(p)}, \tag{B.11b}$$

and the $N_{\pm p}^{\pm(n-1)}$ are given by the following recursions:

$$N_p^{(-p)} = 0, \tag{B.12a}$$

$$N_p^{(p)} = I, \tag{B.12b}$$

$$\left.\begin{aligned} &N_p^{(-i)} = N_p^{(1-i)} - \bar{m}_{-i}Z_{-i}N_p^{(i-1)}, && \text{(B.13a)}\\ &N_p^{(i)} = N_p^{(i-1)} - \bar{m}_iZ_iN_p^{(-i)}, && \text{(B.13b)}\end{aligned}\right\} \quad i = p+1, \dots, n-1,$$

$$N_{-p}^{(-p)} = I, \tag{B.14a}$$

$$N_{-p}^{(p)} = -\bar{m}_pZ_p, \tag{B.14b}$$

$$\left.\begin{aligned} &N_{-p}^{(-i)} = N_{-p}^{(1-i)} - \bar{m}_{-i}Z_{-i}N_{-p}^{(i-1)}, && \text{(B.15a)}\\ &N^{(i)}{}_{-p} = N_{-p}^{(i-1)} - \bar{m}_iZ_iN^{(-i)}_{-p}, && \text{(B.15b)}\end{aligned}\right\} \quad i = p+1, \dots, n-1.$$

Outline of proof. It can be shown by a straightforward induction [11] on k, using the definitions of $\hat{T}^{(\pm k)}$ and $N_{\pm p}^{(\pm k)}$, that

$$\bar{T}^{(-k)} - \hat{T}^{(-k)} = \sum_{p=1}^{k} (N_p^{(-k)}E^{(p)} + N_{-p}^{(-k)}E^{(-p)}), \qquad k = 1, \dots, n-1,$$

$$\bar{T}^{(k)} - \hat{T}^{(k)} = \sum_{p=1}^{k} (N_p^{(k)}E^{(p)} + N_{-p}^{(k)}E^{(-p)}), \qquad k = 1, \dots, n-1.$$

Setting $k = n - 1$ yields (B.8a) and (B.9a). The proof of (B.8b) and (B.9b) is analogous, with references to $E^{(\pm p)}$ being replaced by references to $\mathbf{h}^{(\pm p)}$.

We now evaluate a bound for $|f_{ij}|$. Consider the term $N_p^{(1-n)}E^{(p)}$ in (B.8a). From (B.10b),

$$e_{ij}^{(p)} = \bar{t}_{ij}^{(p)} - (\bar{t}_{ij}^{(p-1)} - \bar{m}_p \bar{t}_{i+p,j}^{(-p)}). \tag{B.16}$$

Assume that the following holds:

$$fl(x \cdot y) = (1+\delta)(x \cdot y), \qquad |\delta| \leqq \mu, \tag{B.17}$$

where $\cdot$ is any of addition, subtraction, multiplication, or division, fl is the floating-point result of this operation, and μ is machine precision. In (B.16), the left-hand side $e_{ij}^{(p)}$ is due to the floating-point rounding errors arising in the term in parentheses on the right-hand side. Using (B.17) and neglecting second-order quantities, we can derive the following expression for these rounding errors:

$$e_{ij}^{(p)} = \delta \bar{t}_{ij}^{(p-1)} - (\delta + \delta')\bar{m}_p \bar{t}_{i+p,j}^{(-p)}, \qquad |\delta|, |\delta'| \leqq \mu. \tag{B.18}$$

By considering the Bareiss recursion (2.5), (2.7), it may be shown by a straightforward induction on p that, neglecting second-order quantities, we get

$$|\bar{t}_{ij}^{(p-1)}| \leqq \tau \prod_{l=1-p}^{p-1} (1 + |m_l|), \tag{B.19}$$

$$|\bar{t}_{ij}^{(-p)}| \leqq \tau \prod_{l=-p}^{p-1} (1 + |m_l|), \tag{B.20}$$

where $m_0 = 0$ and $\tau = \max_{i,j} |t_{ij}|$. Substituting (B.19) and (B.20) into (B.18), we get the bound

$$|e_{ij}^{(p)}| \leqq 2\mu\tau \prod_{l=-p}^{p} (1 + |m_l|). \tag{B.21}$$

Postmultiplying the recursion (B.13) by $E^{(p)}$ we may show, by a straightforward induction on k, that

$$(N_p^{(-k)}E^{(p)})_{ij} \leqq \max_{i,j} |e_{ij}^{(p)}| \prod_{l=p+1}^{k} (1 + m_l)(1 + m_{-l}).$$

Setting $k = n - 1$ and using (B.21) we get

$$(N_p^{(1-n)}E^{(p)})_{ij} \leqq 2\mu\tau \prod_{l=1-n}^{n-1} (1 + |m_l|). \tag{B.22}$$

Similarly, we can bound

$$(N_{-p}^{(1-n)}E^{(-p)})_{ij} \leqq 2\mu\tau \prod_{l=1-n}^{n-1} (1 + |m_l|). \tag{B.23}$$

Putting (B.22) and (B.23) into (B.8a) yields the bound (2.19a) for $|f_{kl}|$. The bounds (2.19b) and (2.20) follow similarly. □

Appendix C. Proof of Lemma 3.1. It is shown by Bareiss [2] that

$$\mathbf{t}_{i\cdot}^{(-j)} = \mathbf{t}_{i\cdot} + \sum_{k=1}^{j} \alpha_k \mathbf{t}_{i-k\cdot}, \qquad i = j+1, \ldots, n \tag{C.1}$$

for some $\{\alpha_1, \dots, \alpha_j\}$. Then, from the first $j+2$ columns of (C.1) for $i = j+2$, we have

$$\text{(C.2)} \qquad \mathbf{t}^{(-j)}_{j+2,1:j+2} = [t^{(-j)}_B, 0, \dots, 0, t^{(-j)}_A] = \mathbf{t}_{j+2,1:j+2} + \sum_{k=1}^{j} \alpha_k \mathbf{t}_{j+2-k,1:j+2}.$$

Define $\mathbf{e}_k \equiv [0, \dots, 0, 1]^T$ for any k. It is clear from (C.2) that

$$\text{(C.3)} \qquad T'_{j+2} \equiv T_{j+2} - \mathbf{e}_{j+2}[t^{(-j)}_B, 0, \dots, 0, t^{(-j)}_A]$$

is singular. But Kahan [7] has shown that

$$\begin{aligned} \frac{1}{\|T^{-1}_{j+2}\|} = \frac{\|T_{j+2}\|}{\kappa(T_{j+2})} &= \min_{T_{j+2}+E \text{ singular}} \|E\| \\ &\leqq \|\mathbf{e}_{j+2}[t^{(-j)}_B, 0, \dots, 0, t^{(-j)}_A]\| \quad \text{using (C.3)} \\ &\leqq |t^{(-j)}_A| + |t^{(-j)}_B|. \end{aligned}$$

□

Appendix D. Proof of Lemma 3.2. This proof is adapted from that of [11, pp. 126–128]. We prove the lemma only for the case when $p > 0$. The proof for $p < 0$ is similar. From (C.1), and noting that $\mathbf{t}^{(-j)}_{i+j\cdot} = \mathbf{t}^{(-j)}_{*i\cdot}$, $i = 1, \dots, n-j$,

$$\text{(D.1)} \qquad \mathbf{t}^{(-j)}_{*i\cdot} = \mathbf{t}_{j+i\cdot} + \sum_{k=1}^{j} \alpha_k \mathbf{t}_{j+i-k\cdot}, \qquad i = 1, \dots, n-j,$$

and it can be shown, in a manner analogous to Bareiss's proof of (C.1), that for any sequence of "normal" Bareiss cycles (2.9)–(2.12) and "pivot" cycles (in which BPP is used to move the ZB up or down; see § 3.5):

$$\text{(D.2)} \qquad \tilde{\mathbf{t}}^{(-j)}_{*i\cdot} = \tilde{\alpha}_0 \mathbf{t}_{j+i\cdot} + \sum_{k=1}^{j} \tilde{\alpha}_k \mathbf{t}_{j+i-k\cdot}, \; i = 1, \dots, n-j, \; \tilde{\alpha}_0, \dots, \tilde{\alpha}_j \text{ not all zero.}$$

From (D.1) and (D.2), we have, respectively,

$$\text{(D.3)} \qquad \mathbf{t}^{(-j)}_{*1,1+p:j+p} = \mathbf{t}_{j+1,1+p:j+p} + \sum_{k=1}^{j} \alpha_k \mathbf{t}_{j+1-k,1+p:j+p},$$

$$\text{(D.4)} \qquad \tilde{\mathbf{t}}^{(-j)}_{*1,1+p:j+p} = \tilde{\alpha}_0 \mathbf{t}_{j+1,1+p:j+p} + \sum_{k=1}^{j} \tilde{\alpha}_k \mathbf{t}_{j+1-k,1+p:j+p}.$$

But if the ZB is at displacement p, then

$$\text{(D.5)} \qquad \tilde{\mathbf{t}}^{(-j)}_{*1,1+p:j+p} = \mathbf{0}^T,$$

and since (by assumption) $T_{j;p}$ is nonsingular, the summation on the right-hand side of (D.4) is not zero, so $\tilde{\alpha}_0 \neq 0$. Substituting (D.5) into (D.4) and dividing by $\tilde{\alpha}_0$, we get

$$\text{(D.6)} \qquad \mathbf{0}^T = \mathbf{t}_{j+1,1+p:j+p} + \sum_{k=1}^{j} \alpha'_k \mathbf{t}_{j+1-k,1+p:j+p},$$

where $\alpha'_k \equiv \tilde{\alpha}_k / \tilde{\alpha}_0$, $k = 1, \dots, j$. Subtracting (D.6) from (D.3), we get

$$\text{(D.7)} \qquad \mathbf{t}^{(-j)}_{*1,1+p:j+p} = [\alpha_j - \alpha'_j, \alpha_{j-1} - \alpha'_{j-1}, \dots, \alpha_1 - \alpha'_1] T_{j;p}.$$

Subtracting (D.1) from $\tilde{\alpha}_0^{-1} \times$ (D.2) and using (D.7), we get for column m,

$$\tilde{\alpha}_0^{-1} \tilde{t}^{(-j)}_{*im} - t^{(-j)}_{*im} = -\mathbf{t}^{(-j)}_{*1,1+p:j+p} T^{-1}_{j;p} \mathbf{t}_{i:i+j-1,m}, \qquad i = 1, \dots, n-j.$$

Using the "diagonal" notation (§ 3.3) for $T_*^{(-j)}$ and $\tilde{T}_*^{(-j)}$, and noting that $t_{*1m}^{(-j)} = 0$, $1 \leqq m \leqq j$, we can write this as

$$\alpha_0^{-1}\tilde{t}_{k-p}^{(-j)} - \dot{t}_k^{(-j)} = -(\mathbf{0}^T, t_1^{(-j)}, \ldots, t_p^{(-j)}) T_{j;p}^{-1}\mathbf{t}_{i:i+j-1,m},$$

$$k = \begin{cases} m-i-j+1, & m \geqq i+p+j, \\ m-i, & m < i+p, \end{cases}$$

where

$$\dot{t}_k^{(-j)} = \begin{cases} t_k^{(-j)}, & k = j+1-n, \ldots, -1, \quad p+1, \ldots, n-j, \\ 0, & k = 0, \ldots, p-1. \end{cases}$$

Therefore,

$$\alpha_0^{-1}\tilde{t}_{k-p}^{(-j)} = \dot{t}_k^{(-j)} + \varepsilon_k, \tag{D.8}$$

where

$$\varepsilon_k = -(\mathbf{0}^T, t_1^{(-j)}, \ldots, t_p^{(-j)}) T_{j;p}^{-1}\mathbf{t}_{i:i+j-1,m}.$$

Taking norms, we bound $|\varepsilon_k|$ by

$$|\varepsilon_k| \leqq \beta_p \kappa(T_{j;p}) \|\mathbf{0}^T, t_1^{(-j)}, \ldots, t_p^{(-j)}\|, \tag{D.9}$$

where

$$\beta_p = \|T\| / \|T_{j;p}\|.$$

The result follows from (D.8) and (D.9). □

Appendix E. Derivation of steps 7 and 8 of Algorithm 3.2. Define

$$\begin{aligned} &\text{(E.1a)} \quad \hat{T}'^{\pm(j-i)}_{*} \equiv \hat{T}^{\pm(j-i)}_{*} / m^{\pm(j-i)}_{*}, \\ &\text{(E.1b)} \quad \hat{M}'^{\pm(j-i)}_{*} \equiv \hat{M}^{\pm(j-i)}_{*} / m^{\pm(j-1)}_{*}, \end{aligned} \qquad i = 1, 2.$$

Then, for the Bareiss algorithm, it can be verified from (2.9) and (2.10) that

$$\hat{T}'^{(1-j)}_{*} = \hat{T}'^{(2-j)}_{*F1} - \frac{\hat{t}'^{(2-j)}_B}{\hat{t}'^{(j-2)}_B} \hat{T}'^{(j-2)}_{*L1}. \tag{E.2}$$

Substituting (E.1a) and (E.1b) into (E.2) and rearranging yields

$$\alpha \hat{T}'^{(j-2)}_{*L1} = \hat{T}^{(2-j)}_{*F1} - \frac{\hat{m}^{(2-j)}_*}{\hat{m}^{(1-j)}_*} \hat{T}^{(1-j)}_*, \tag{E.3}$$

where

$$\alpha = \hat{m}^{(2-j)}_* \hat{t}'^{(2-j)}_B / \hat{t}'^{(j-2)}_B.$$

Similarly,

$$\alpha \hat{M}'^{(j-2)}_{*L1} = \hat{M}^{(2-j)}_{*F1} - \frac{\hat{m}^{(2-j)}_*}{\hat{m}^{(1-j)}_*} \hat{M}^{(1-j)}_*. \tag{E.4}$$

We can scale $\hat{T}^{(j-2)}_{*L1}$ and $\hat{M}^{(j-2)}_{*L1}$ arbitrarily, provided that $\hat{T}^{(j-2)}_{*L1} = \hat{M}^{(j-2)}_{*L1} T$. This is satisfied if we set

$$\hat{T}^{(j-2)}_{*L1} = \alpha \hat{T}'^{(j-2)}_{*L1} \quad \text{and} \quad \hat{M}^{(j-2)}_{*L1} = \alpha \hat{M}'^{(j-2)}_{*L1}, \tag{E.5}$$

since by definition $\hat{T}'^{(j-2)}_{*L1} = \hat{M}'^{(j-2)}_{*L1}T$. Steps 7 and 8 of Algorithm 3.2 then follow from (E.3), (E.4), and (E.5). □

Appendix F. Proof of Theorem 4.1. The proof is adapted from that of [11, pp. 157–159]. We first note that the recursions for $m_{UT}^{(\pm j)}$ and $m_{LT}^{(\pm j)}$ are easily verified using the multiplier matrix recursions ((3.5), (3.6) for B-cycles; (5.9) for general form).

We now show that L is unit lower triangular and U is upper triangular. Let g_j and h_j be the values of g and h, respectively, after cycle j of Algorithm 4.1, step 5, and let $n_U(j)$ and $n_D(j)$ be the number of U- and D-cycles, respectively, after cycle j of PBA. It is easily seen from loops 1.2 and 5 that

$$\text{(F.1)} \quad g_j = n_U - n_U(j) \quad \text{and} \quad s_{\min}(j+1) \equiv \min\{s_i\}_1^{j+1} = n_U + 1 - n_U(j),$$

$$\text{(F.2)} \quad h_j = n_D - n_D(j) \quad \text{and} \quad v_{\min}(j+1) \equiv \min\{v_i\}_1^{j+1} = n_D + 1 - n_D(j).$$

It is also clear from loop 1.2 that

$$\mathbf{s}_j \equiv [s_1, \dots, s_j]^T = P^{(j)}[s_{\min}(j), \dots, s_{\min}(j) + j - 1]^T,$$

where $P^{(j)}$ is an order-j permutation matrix. With (F.1), this becomes

$$\text{(F.3)} \qquad \mathbf{s}_j = P^{(j)}[g_{j-1} + 1, \dots, g_{j-1} + j]^T.$$

Now, by definition,

$$\{t_{b(g_{j-1}),i}^{(\pm j)}\}_{i=n-g_{j-1}}^{n-g_{j-1}-j+1}$$

are all zero, so

$$\{(\mathbf{t}_{b(g_{j-1})\cdot}^{(\pm j)RT})_i\}_{i=g_{j-1}+1}^{g_{j-1}+j}$$

are all zero. By (4.7a), premultiplication by P moves rows $s_1, \dots, s_n$ to rows $1, \dots, n$. Hence, by (F.3), rows $g_{j-1}+1$ to $g_{j-1}+j$ are moved to positions 1 to j, so

$$\{(P\mathbf{t}_{b(g_{j-1})\cdot}^{(\pm j)})_i\}_{i=1}^{j}$$

are all zero, which implies from (4.10e), (4.11e), and (4.12e) that L is lower triangular. This fact, together with (4.10f), (4.11f) and (4.12f), implies that L is unit lower triangular. The proof that U is upper triangular is similar to the proof that L is lower triangular.

We now show that $LU = PTQ^T$. Let d_j be the displacement of the ZB from its Bareiss position at step j. We use the convention that d_j is positive for an upward displacement and negative for a downward one. Recall from Fig. 3.7(a) that the Toeplitz parts of $T^{(-j)}$ and $T^{(j)}$ have the forms (for $d_j < 0$) shown in Fig. F.1.

Recall from the definition and form of $M^{(\pm j)}$ (see (3.4) and Fig. 3.8) that

$$\text{(F.4)} \qquad \mathbf{t}_{*i\cdot}^{(-j)} = \sum_{l=1}^{j+1} m_{*l}^{(-j)}\mathbf{t}_{l+i-1\cdot}, \qquad \mathbf{t}_{*i\cdot}^{(j)} = \sum_{l=1}^{j+1} m_{*l}^{(j)}\mathbf{t}_{l+i-1\cdot},$$

where $m_{*l}^{(\pm j)} \equiv m_{*1,l}^{(\pm j)}$. This fact, together with the shapes of the $T_*^{(-j)}$ and $T_*^{(j)}$, implies that

$$\text{(F.5)} \quad \mathbf{t}_{a(h_{j-1})\cdot}^{(-j)} = \sum_{l=1}^{j+1} m_{*l}^{(-j)}\mathbf{t}_{h_{j-1}-d_j+l\cdot} \quad \text{and} \quad \mathbf{t}_{a(h_{j-1})\cdot}^{(j)} = \sum_{l=1}^{j+1} m_{*l}^{(j)}\mathbf{t}_{h_{j-1}-d_j+l-1\cdot}.$$

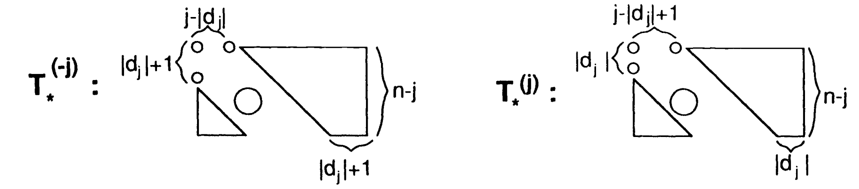


FIG. F.1. *Forms of* $T_*^{(-j)}$ *and* $T_*^{(j)}$.

Since d_j increments with U-cycles and decrements with D-cycles, we must have

$$d_j = d_0 + n_U(j) - n_D(j) = n_D - n_U + n_U(j) - n_D(j).$$

Therefore, if cycle j was a U- or B-cycle, it can be shown from (F.1) and (F.2) that $h_{j-1} - d_j + 1 = s_{\min}(j+1)$, so (F.5) can be written

$$\text{(F.6)} \qquad \begin{aligned} \mathbf{t}^{(-j)}_{a(h_{j-1})\cdot} &= m^{(-j)}_{*1}\mathbf{t}_{s_{\min}(j+1)\cdot} + \cdots + m^{(-j)}_{*j+1}\mathbf{t}_{s_{\max}(j+1)\cdot} \\ &= m^{(-j)}_{*s_1-\sigma}\mathbf{t}_{s_1\cdot} + m^{(-j)}_{*s_2-\sigma}\mathbf{t}_{s_2\cdot} + \cdots + m^{(-j)}_{*s_{j+1}-\sigma}\mathbf{t}_{s_{j+1}\cdot} \end{aligned}$$

by rearrangement, where $\sigma \equiv s_{\min}(j+1) - 1$. Now for a B-cycle, $s_{j+1} = s_{\min}(j+1) + j$, so $m^{(-j)}_{*s_{j+1}-\sigma} = m^{(-j)}_{*j+1} = m^{(-j)}_{UT}$. Similarly, for a U-cycle, $m^{(-j)}_{s_{j+1}-\sigma} = m^{(-j)}_{LT}$. So, for both B- and U-cycles, by (F.6), (4.10g), and (4.11g), we have

$$\mathbf{u}_{j+1\cdot}Q = [m^{(-j)}_{*s_1-\sigma}, \ldots, m^{(-j)}_{*s_j-\sigma}, m^{(-j)}_{ST}][\mathbf{t}^T_{s_1\cdot}, \ldots, \mathbf{t}^T_{s_{j+1}\cdot}]^T / m^{(-j)}_{ST},$$

where $m^{(-j)}_{ST} = m^{(-j)}_{UT}$ for a B-cycle and $m^{(-j)}_{ST} = m^{(-j)}_{LT}$ for a U-cycle. By the definition of P, this becomes

$$\mathbf{u}_{j+1\cdot}Q = [m^{(-j)}_{*s_1-\sigma}, \ldots, m^{(-j)}_{*s_j-\sigma}, m^{(-j)}_{ST}](PT)_{1:j+1\cdot} / m^{(-j)}_{ST}.$$

Since $\mathbf{u}_{j+1\cdot}Q$ is a normalized linear combination of the first $j+1$ rows of PT, $\mathbf{u}_{j+1\cdot}$ is row j of the unique U-factor of PTQ^T. The proof when cycle j is a D-cycle is similar. It is also clear from (4.9c) and (4.3a) that $\mathbf{u}_{1\cdot}Q$ is row s_1 of T, i.e., row 1 of PT, so $\mathbf{u}_{1\cdot}$ is row 1 of the U-factor of PTQ^T.

We can similarly show for any j that $\mathbf{l}_{\cdot j+1}$ is column j of the unique L-factor of PTQ^T. □

REFERENCES

[1] G. A. Baker, *Essentials of Padé Approximants*, Academic Press, New York, 1975.

[2] E. H. Bareiss, *Numerical solution of linear equation with Toeplitz and vector Toeplitz matrices*, Numer. Math., 13 (1969), pp. 404–424.

[3] J. R. Bunch, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comp., 6 (1985), pp. 349–364.

[4] G. Cybenko, *The numerical stability of the Levinson-Durbin algorithm for Toeplitz systems of equations*, SIAM J. Sci. Statist. Comp., 1 (1980), pp. 303–310.

[5] P. Delsarte, Y. V. Genin, and Y. G. Kamp, *A generalization of the Levinson algorithm for Hermitian Toeplitz matrices with any rank profile*, IEEE Trans. Acoust. Speech Signal Process., ASSP-33 (1985), pp. 964–971.

[6] G. Heinig, *Inversion of Toeplitz and Hankel matrices with singular sections*, Wiss. Z. d. Techn. Hochsch. Karl-Marx-Stadt, 25 (1983), pp. 326–333.

[7] W. Kahan, *Numerical linear algebra*, Canad. Math. Bull., 9 (1966), pp. 757–801.

[8] S. Y. Kung, *A Toeplitz approximation method and some applications*, in Proc. 1981 Internat. Symp. Math. Theory of Networks and Systems, Santa Monica, CA, 1981, pp. 262–266.

[9] N. Levinson, *The Wiener* RMS (*root-mean-square*) *error criterion in filter design and prediction*, J. Math. Phys., 25 (1946), pp. 261–278.

[10] J. Makhoul, *On the eigenvectors of symmetric Toeplitz matrices*, IEEE Trans. Acoust. Speech Signal Process., ASSP-29 (1981), pp. 868–872.

[11] D. R. Sweet, *Numerical methods for Toeplitz matrices*, Ph.D. thesis, Adelaide University, Adelaide, Australia, 1982.

[12] ———, *The use of pivoting to improve the numerical performance of Toeplitz solvers*, Proc. SPIE, 696 (1987), pp. 8–18.

[13] ———, *The propagation of rounding errors in pivoting techniques for Toeplitz matrix solvers*, in Proc. Internat. Conf. on Computational Techniques and Applications (CTAC-89), W. L. Hogarth and B. J. Noye, eds., Hemisphere, New York, Washington, Philadelphia, London, 1990, pp. 623–630.

[14] W. F. Trench, *An algorithm for the inversion of finite Toeplitz matrices*, SIAM J. Appl. Math., 12 (1964), pp. 515–522.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 494–499, April 1993

UPDATING A RANK-REVEALING ULV DECOMPOSITION*

G. W. STEWART[†]

Abstract. A ULV decomposition of a matrix A of order n is a decomposition of the form $A = ULV^{\mathrm{H}}$, where U and V are orthogonal matrices and L is a lower triangular matrix. When A is approximately of rank k, the decomposition is rank revealing if the last $n-k$ rows of L are small. This paper presents algorithms for updating a rank-revealing ULV decomposition. The algorithms run in $O(n^2)$ time, and can be implemented on a linear array of processors to run in $O(n)$ time.

Key words. ULV decomposition, URV decomposition, rank revealing, updating

AMS(MOS) subject classifications. 65F25, 65F25, 65F35

1. Introduction.

Let A be an $m \times n$ matrix. A rank-revealing URV decomposition of A is a reduction of A by unitary transformations to a triangular matrix of the form

$$U^{\mathrm{H}}AV = \begin{pmatrix} R & H \\ 0 & E \\ 0 & 0 \end{pmatrix}. \tag{1.1}$$

The decomposition is rank revealing in the sense that the matrices H and E are smaller than some prespecified tolerance, and the smallest singular value of R is greater than that tolerance. Such decompositions—they are not unique—are useful in solving rank-deficient systems. Moreover, if $V = (V_1\ V_2)$ is partitioned conformally, then in the spectral norm

$$\|AV_2\| = \left\| \begin{pmatrix} H \\ E \end{pmatrix} \right\|, \tag{1.2}$$

so that the columns of V_2 form an orthonormal basis for an approximate null space of A, something required in signal processing applications like direction of arrival estimation.

The advantage of the URV decomposition over the more familiar singular value decomposition is that it can be updated when a row is added to A (in many applications, A is first multiplied by a constant less than one to damp out old data, a process known as exponential windowing). The updating procedure, which is described in [7], requires $O(n^2)$ operations and preserves the rank-revealing character of the decomposition. Moreover, it can be implemented on a linear array of n processors to run in $O(n)$ time.

Although the URV decomposition is fully satisfactory for applications like recursive least-squares, it is less satisfactory for applications in which an approximate null space is needed. The reason is the presence of H in (1.2). To see that it should not be there, let $(U_1\ U_2)$ be a partition of U conformal with (1.1). Then it is easily seen that $\|U_2^{\mathrm{H}}A\| = \|E\|$. Consequently the last $n-k$ singular values of A are less than or equal to $\|E\|$, and the corresponding left singular vectors form an approximate null

* Received by the editors April 17, 1991; accepted for publication (in revised form) August 23, 1991.

[†] Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland 20742 (stewart@cs.umd.edu). This work was supported in part by Air Force Office of Scientific Research contract AFOSR-87-0188.

$$\begin{array}{cccccccccc}
\rightarrow & \check{w} & & & 0 & & & 0 & & 0\\
\rightarrow & w & \Longrightarrow & \rightarrow & \check{w} & \Longrightarrow & & 0 & \Longrightarrow & 0\\
 & w & & \rightarrow & w & & \rightarrow & \check{w} & & 0\\
 & w & & & w & & \rightarrow & w & & 1
\end{array}$$

FIG. 2.1. *Reduction of W.*

space whose residual has norm less than or equal to $\|E\|$. Thus V_2 is not the best available approximate null space.

In [7] this problem was circumvented by including a refinement step that reduces the size of H. The properties of this refinement have been analyzed in [8]. In experiments with the MUSIC algorithm for direction of arrival estimation, the refinement was found to improve the results [1]. However, it adds extra work, and aesthetically it has the appearance of a stopgap. The purpose of this paper is to present an alternative.

Specifically, we will describe how to update a lower triangular decomposition of the form

$$U^{\mathrm{H}}AV = \begin{pmatrix} L & 0 \\ H & E \\ 0 & 0 \end{pmatrix}, \tag{1.3}$$

where U and V are orthogonal, L is well conditioned, and H and E are small. We will call such a decomposition a rank-revealing ULV decomposition. The updating algorithm consists of two parts: an algorithm to bring a lower triangular matrix into rank-revealing ULV form and the updating algorithm proper. We will present the former in the next section and the latter in §3. The paper concludes with some general observations on the algorithms.

2. Deflation.

Although adding a row to a matrix cannot decrease its rank, in many applications the matrix is first multiplied by a constant less than one to damp out old information. Under such circumstances, it is possible for the matrix L in (1.3) to become effectively rank deficient. Here we present an algorithm to calculate a rank-revealing ULV decomposition of a lower triangular matrix L.

The first step is to determine a vector w of norm one such that $\omega = \|w^{\mathrm{H}}L\|$ approximates the smallest singular value of L. This can be done by using any of a number of reliable condition estimators [4].[1] If ω is greater than a prescribed tolerance, then there is nothing to be done. Otherwise, we must modify L by unitary transformations so that its last row becomes small. We will use plane rotations to accomplish this reduction. The reader is assumed to be familiar with plane rotations, which are discussed in most texts on numerical linear algebra (e.g., see [3], [6], and [9]).

We begin by reducing w the nth unit vector $\mathbf{e}_n$. The reduction is illustrated in Fig. 2.1. The two arrows represent the plane of the rotation, and it annihilates the component of w with a check over it. Denote the product of rotations by

$$P = P_1P_2\cdots P_{n-1}.$$

[1] It is worth noting that these estimators succeed, where QR with pivoting fails, in revealing the rank of a widely cited matrix of Kahan [5].

$$
\begin{array}{ccccc}
 & & & & \\
\rightarrow & l & 0 & 0 & 0 \\
\rightarrow & l & l & 0 & 0 \\
 & l & l & l & 0 \\
 & l & l & l & l
\end{array}
\Longrightarrow
\begin{array}{cccc}
\downarrow & \downarrow & & \\
l & \check{l} & 0 & 0 \\
l & l & 0 & 0 \\
l & l & l & 0 \\
l & l & l & l
\end{array}
\Longrightarrow
\begin{array}{ccccc}
 & & & & \\
 & l & 0 & 0 & 0 \\
\rightarrow & l & l & 0 & 0 \\
\rightarrow & l & l & l & 0 \\
 & l & l & l & l
\end{array}
\Longrightarrow
$$

$$
\begin{array}{cccc}
 & \downarrow & \downarrow & \\
l & 0 & 0 & 0 \\
l & l & \check{l} & 0 \\
l & l & l & 0 \\
l & l & l & l
\end{array}
\Longrightarrow
\begin{array}{ccccc}
 & & & & \\
 & l & 0 & 0 & 0 \\
 & l & l & 0 & 0 \\
\rightarrow & l & l & l & 0 \\
\rightarrow & l & l & l & l
\end{array}
\Longrightarrow
\begin{array}{cccc}
 & & \downarrow & \downarrow \\
l & 0 & 0 & 0 \\
l & l & 0 & 0 \\
l & l & l & \check{l} \\
l & l & l & l
\end{array}
\Longrightarrow
\begin{array}{cccc}
 & & & \\
l & 0 & 0 & 0 \\
l & l & 0 & 0 \\
l & l & l & 0 \\
h & h & h & e
\end{array}
$$

FIG. 2.2. *Reduction of PL.*

Next we apply the rotations P_i to L from the left, as is shown in Fig. 2.2. The application of P_i produces a nonzero element in the $(i, i+1)$-element of L. This element is removed by postmultiplying by a plane rotation Q_i. Let $Q = Q_1 Q_2 \cdots Q_{n-1}$ be the product of these rotations.

The appearance of h's and e in the last matrix of Fig. 2.2 is meant to indicate that the last row of PLQ is small. To see that this is true, write

$$\omega = \|w^{\mathrm{H}} L\| = \|w^{\mathrm{H}} P^{\mathrm{H}} PLQ\| = \|\mathbf{e}_n^{\mathrm{T}}(PLQ)\|.$$

Thus the last row of PLQ has norm ω.

If the $(n-1) \times (n-1)$ leading principle submatrix of L is sufficiently well conditioned, we have our rank-revealing ULV decomposition. If not, we can repeat the deflation procedure until a sufficiently well conditioned matrix is found in the upper right-hand corner.

3. Updating. We now turn to the updating step. Specifically, we suppose that we are given an additional row z^{H} and wish to determine a rank-revealing ULV decomposition of

$$\begin{pmatrix} A \\ z^{\mathrm{H}} \end{pmatrix}.$$

We begin by forming $z^{\mathrm{H}} V$ to bring the row into the coordinate system of the current decomposition. Thus the problem of updating reduces to the problem of updating

$$\begin{pmatrix} L & 0 \\ H & E \\ x^{\mathrm{H}} & y^{\mathrm{H}} \end{pmatrix},$$

where we have partitioned $z^{\mathrm{H}} V = (x^{\mathrm{H}} \; y^{\mathrm{H}})$.

We now reduce y^{H} to $\|y\| \mathbf{e}_1^{\mathrm{T}}$, as is shown in Fig. 3.1, where only the parts cor-

$$
\begin{array}{cccc}
 & & \downarrow & \downarrow \\
e & 0 & 0 & 0 \\
e & e & 0 & 0 \\
e & e & e & 0 \\
e & e & e & e \\
y & y & y & \check{y}
\end{array}
\Longrightarrow
\begin{array}{ccccc}
 & e & 0 & 0 & 0 \\
 & e & e & 0 & 0 \\
\rightarrow & e & e & e & \check{e} \\
\rightarrow & e & e & e & e \\
 & y & y & y & 0
\end{array}
\Longrightarrow
\begin{array}{cccc}
 & \downarrow & \downarrow & \\
e & 0 & 0 & 0 \\
e & e & 0 & 0 \\
e & e & e & 0 \\
e & e & e & e \\
y & y & \check{y} & 0
\end{array}
\Longrightarrow
$$

$$
\begin{array}{ccccc}
 & e & 0 & 0 & 0 \\
\rightarrow & e & e & \check{e} & 0 \\
\rightarrow & e & e & e & 0 \\
 & e & e & e & e \\
 & y & y & 0 & 0
\end{array}
\Longrightarrow
\begin{array}{cccc}
\downarrow & \downarrow & & \\
e & 0 & 0 & 0 \\
e & e & 0 & 0 \\
e & e & e & 0 \\
e & e & e & e \\
y & \check{y} & 0 & 0
\end{array}
\Longrightarrow
\begin{array}{ccccc}
\rightarrow & e & \check{e} & 0 & 0 \\
\rightarrow & e & e & 0 & 0 \\
 & e & e & e & 0 \\
 & e & e & e & e \\
 & y & 0 & 0 & 0
\end{array}
\Longrightarrow
\begin{array}{cccc}
e & 0 & 0 & 0 \\
e & e & 0 & 0 \\
e & e & e & 0 \\
e & e & e & e \\
y & 0 & 0 & 0
\end{array}
$$

FIG. 3.1. *Reducing y.*

$$
\begin{array}{cccccc}
 & l & 0 & 0 & 0 & 0 \\
 & l & l & 0 & 0 & 0 \\
 & l & l & l & 0 & 0 \\
\rightarrow & h & h & h & e & 0 \\
 & h & h & h & e & e \\
\rightarrow & x & x & x & \check{y} & 0
\end{array}
\Longrightarrow
\begin{array}{cccccc}
 & l & 0 & 0 & 0 & 0 \\
 & l & l & 0 & 0 & 0 \\
\rightarrow & l & l & l & 0 & 0 \\
 & x & x & x & y & 0 \\
 & h & h & h & e & e \\
\rightarrow & x & x & \check{x} & 0 & 0
\end{array}
\Longrightarrow
$$

$$
\begin{array}{cccccc}
 & l & 0 & 0 & 0 & 0 \\
\rightarrow & l & l & 0 & 0 & 0 \\
 & l & l & l & 0 & 0 \\
 & x & x & x & y & 0 \\
 & h & h & h & e & e \\
\rightarrow & x & \check{x} & 0 & 0 & 0
\end{array}
\Longrightarrow
\begin{array}{cccccc}
\rightarrow & l & 0 & 0 & 0 & 0 \\
 & l & l & 0 & 0 & 0 \\
 & l & l & l & 0 & 0 \\
 & x & x & x & y & 0 \\
 & h & h & h & e & e \\
\rightarrow & \check{x} & 0 & 0 & 0 & 0
\end{array}
\Longrightarrow
\begin{array}{ccccc}
l & 0 & 0 & 0 & 0 \\
l & l & 0 & 0 & 0 \\
l & l & l & 0 & 0 \\
x & x & x & y & 0 \\
h & h & h & e & e \\
0 & 0 & 0 & 0 & 0
\end{array}
$$

FIG. 3.2. *Triangularization step.*

responding to y^{H} and E are shown. At the end of the reduction the matrix has the form

$$
\begin{array}{ccccc}
l & 0 & 0 & 0 & 0 \\
l & l & 0 & 0 & 0 \\
l & l & l & 0 & 0 \\
h & h & h & e & 0 \\
h & h & h & e & e \\
x & x & x & y & 0
\end{array}
$$

One way to finish the update is to continue as in Fig. 3.2 to incorporate x and what is left of y into the decomposition. This process moves the presumably large elements of x^{H} into the first row of H. If there has been no increase in rank, this destroys the rank-revealing character of the matrix. However, in that case the deflation procedure

$$\begin{array}{ccccc} \downarrow & & & \downarrow & \\ l & 0 & 0 & \check{y} & 0 \\ l & l & 0 & y & 0 \\ l & l & l & y & 0 \\ h & h & h & e & 0 \\ h & h & h & e & e \\ 0 & 0 & 0 & y & 0 \end{array} \Longrightarrow \begin{array}{ccccc} & \downarrow & & \downarrow & \\ l & 0 & 0 & 0 & 0 \\ l & l & 0 & \check{y} & 0 \\ l & l & l & y & 0 \\ h & h & h & e & 0 \\ h & h & h & e & e \\ y & 0 & 0 & y & 0 \end{array} \Longrightarrow \begin{array}{ccccc} & & \downarrow & \downarrow & \\ l & 0 & 0 & 0 & 0 \\ l & l & 0 & 0 & 0 \\ l & l & l & \check{y} & 0 \\ h & h & h & e & 0 \\ h & h & h & e & e \\ y & y & 0 & y & 0 \end{array} \Longrightarrow \begin{array}{ccccc} & & & & \\ l & 0 & 0 & 0 & 0 \\ l & l & 0 & 0 & 0 \\ l & l & l & 0 & 0 \\ h & h & h & e & 0 \\ h & h & h & e & e \\ y & y & y & y & 0 \end{array}$$

FIG. 3.3. *Eliminating the tower of y's.*

will restore the small elements.

If y is small enough, then the rank cannot increase. In that case there is another way of proceeding. Perform the reduction of Fig. 3.2 but skip the first step. This will give a matrix having the form of the first matrix in Fig. 3.3 with a tower of contributions from the scalar y at the bottom. Rotations are used as shown in the figure to reduce the tower. This fills up the bottom row again, but now the elements are of the same size as y. Since y is small, the algorithm of Fig. 3.2 can be used to complete the decomposition without destroying the rank-revealing structure.

It is worth noting that if the y's in the tower are small compared with the diagonal elements of L, the y's along the last row will actually be of order $\|y\|^2$. If only an approximate decomposition is required, it may be possible to neglect them.

4. Comments. We mentioned in the introduction that a ULV decomposition can be expected to give a higher quality approximate null space than a URV decomposition. However, there are trade-offs. It costs more to deflate the URV decomposition if we insist on refinement steps. On the other hand the updating algorithm for the URV decomposition is much simpler. In fact, when there is no change of rank, it amounts to the usual LINPACK updating algorithm SCHUD [2]. Only experience with real-life problems will tell us under what circumstances one decomposition is to be preferred to the other.

Both sets of algorithms are stable and reliable. They are stable because they use orthogonal transformations straightforwardly with no additional implicit relations. They are as reliable as their underlying condition estimators.

In [7] we showed that the algorithms for the URV decomposition could, in principle, be parallelized on a linear array of processors. The same is true of the algorithms for updating a ULV decomposition. Since the techniques to show that the algorithms have parallel implementations are the same for both decompositions, we do not give the details here.

REFERENCES

[1] G. ADAMS, M. F. GRIFFIN, AND G. W. STEWART, *Direction-of-arrival estimation using the rank-revealing* URV *decomposition*, in Proc. IEEE Internat. Conf. Acoustics, Speech, and Signal Processing, Washington, DC, 1991, to appear.

[2] J. J. DONGARRA, J. R. BUNCH, C. B. MOLER, AND G. W. STEWART, LINPACK *User's Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[3] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[4] N. J. HIGHAM, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596.

[5] W. KAHAN, *Numerical linear algebra*, Canad. Math. Bull., 9 (1966), pp. 757–801.

[6] G. W. STEWART, *Modifying pivot elements in Gaussian elimination*, Math. Comp., 28 (1974), p. 1974.

[7] ———, *An updating algorithm for subspace tracking*, Tech. Rep. CS-TR 2494, Dept. of Computer Science, Univ. of Maryland, College Park, MD, 1990; IEEE Trans. Signal Processing, 40 (1992), pp. 1535–1541.

[8] ———, *On an algorithm for refining a rank-revealing* URV *decomposition and a perturbation theorem for singular values*, Tech. Rep. CS-TR 2626, Dept. of Computer Science, Univ. of Maryland, College Park, MD, 1991.

[9] D. S. WATKINS, *Fundamentals of Matrix Computations*, John Wiley, New York, 1991.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 500–507, April 1993

013

ON A MATRIX ALGEBRA RELATED TO THE DISCRETE HARTLEY TRANSFORM*

DARIO BINI† AND PAOLA FAVATI‡

Abstract. A new matrix algebra $\mathcal{H}$, including the set of real symmetric circulant matrices, is introduced. It is proved that all the matrices of $\mathcal{H}$ can be simultaneously diagonalized by the similarity transformation associated to the discrete Hartley transform. An application of this result to the solution of Toeplitz systems by means of the preconditioned conjugate gradient method is presented.

Key words. Hartley transform, Toeplitz matrix, circulant matrix, preconditioned conjugate gradient method

AMS(MOS) subject classifications. 65F10, 65F15

1. Introduction. The discrete Fourier and sine transforms with the fast algorithms for their computation have an important role in many applications. The Fourier matrix and the sine matrix, associated with the analogous discrete transforms, are closely related to two specific matrix algebras: the circulant class [16] and the $\mathcal{T}$ class [4], respectively. In this paper we consider the discrete Hartley transform, introduce the associated matrix, and define a new matrix algebra related to this transform. Then we show an application of this new class to solving Toeplitz systems by means of the preconditioned conjugate gradient method. More precisely, let

$$S = \left(\sin\frac{2\pi ij}{n}\right),$$
$$\hat{S} = \left(\sqrt{\frac{2}{n+1}}\sin\frac{\pi(i+1)(j+1)}{n+1}\right),$$
$$C = \left(\cos\frac{2\pi ij}{n}\right), \qquad i,j = 0,\dots,n-1$$

be $n \times n$ matrices, $\mathbf{i}$ the complex unit such that $\mathbf{i}^2 = -1$, and $A = \mathrm{Circ}(a_0,\dots,a_{n-1})$ the circulant matrix whose first row is $(a_0,\dots,a_{n-1})$, i.e., a matrix where the (i,j)-entry is given by $a_{k(i,j)}$, $k(i,j) = j - i \bmod n$. The structure of a circulant matrix is displayed below:

$$A = \begin{pmatrix} a_0 & a_1 & \dots & a_{n-1} \\ a_{n-1} & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ a_1 & \dots & a_{n-1} & a_0 \end{pmatrix}.$$

We denote by $\mathcal{C}_n$ the set of all the $n \times n$ circulant matrices having entries in the complex field $\mathbf{C}$. The set $\mathcal{C}_n$ is an n-dimensional linear space, closed under the row by column multiplication. That is, $\mathcal{C}_n$ is a matrix algebra.

*Received by the editors December 26, 1990; accepted for publication (in revised form) August 28, 1991. This work was supported by MURST 60 percent and 40 percent funds.

†Dipartimento di Matematica, Università di Pisa, Via Buonarrotti 2, 56100 Pisa, Italy (bini@dm.unipi.it).

‡Istituto di Elaborazione dell'Informazione del C.N.R, Pisa, Italy (favati@vm.iei.pi.cnr.it).

It is well known (see, for instance, [16]) that the *Fourier matrix* $F = 1/\sqrt{n}(C + \mathbf{i}S)$, associated with the discrete Fourier transform (DFT) $\mathbf{x} \rightarrow F\mathbf{x}$, $\mathbf{x} \in \mathbf{C}^n$, is a unitary matrix that diagonalizes, by a similarity transformation, all the circulant matrices. In fact, the following relations hold:

$$\begin{aligned} &F^H F = FF^H = I, \\ (1.1)\quad &F^H AF = \operatorname{diag}(u_0, \ldots, u_{n-1}), \quad \mathbf{u} = (u_i) = \sqrt{n}F^H A\mathbf{e}_1, \quad \mathbf{e}_1 = (1, 0, \ldots, 0)^T, \\ &A \in \mathcal{C}_n. \end{aligned}$$

Similar properties hold for the matrix $\hat{S}$ associated to the sine transform $\mathbf{x} \rightarrow \hat{S}\mathbf{x}$, and for the algebra $\mathcal{T}_n$ of matrices that can be expressed as a polynomial in $Z = (z_{i,j})$, $z_{i,j} = 1$ if $|i - j| = 1$, $z_{i,j} = 0$, otherwise [4]:

$$(1.2)\qquad \begin{aligned} &\hat{S}\hat{S}^T = \hat{S}^2 = I, \\ &\hat{S}B\hat{S} = \operatorname{diag}(v_0, \ldots, v_{n-1}), \quad v_i = \frac{w_i}{\theta_i}, \quad (w_i) = \hat{S}B\mathbf{e}_1, \\ &\theta_i = \sqrt{\frac{2}{n+1}} \sin\frac{\pi(i+1)}{n+1}, \\ &B \in \mathcal{T}_n. \end{aligned}$$

Due to the computational efficiency of the well-known fast algorithms for the discrete Fourier transform and the sine transform computation [21], (1.1) and (1.2) can be used for solving circulant and $\mathcal{T}_n$ systems with a low computational cost (typically $O(n \log n)$ arithmetic operations). This fact has been used for devising fast algorithms for several matrix computations. For instance, in [17] circulant matrices are used in the solution of Toeplitz banded systems, and in [4] and [2], matrices in the class $\mathcal{T}_n$ are used for the numerical computation of the eigenvalues of banded Toeplitz matrices. More recently (see [22], [12]–[15], [19]) circulant matrices have been used as preconditioners in the solution of Toeplitz systems by means of the conjugate gradient method. The result of [22], [15], and [12] have been extended to the class $\mathcal{T}_n$ improving the rate of convergence and reducing the cost per iteration [6]. The key fact on which these results are based is that, given a Toeplitz matrix T, the condition number of the matrix $S^{-1}T$ can be substantially reduced by choosing a suitable matrix S in the class of real symmetric circulant matrices [22], [15], [12] or in the class $\mathcal{T}_n$ [6]. A more systematic analysis of Toeplitz preconditioners is performed in [20]. Other applications of the classes $\mathcal{C}_n$ and $\mathcal{T}_n$ to the evaluation of the complexity of bilinear forms and to devising inversion formulae for Toeplitz matrices are shown in [5], [7], [1], and [3].

In this paper, we introduce the *Hartley matrix* $H = 1/\sqrt{n}(C + S)$ associated to the discrete Hartley transform (DHT) $\mathbf{x} \rightarrow H\mathbf{x}$ (see [8] and [9]), and define the algebra $\mathcal{H}_n$ of all the $n \times n$ real matrices that are simultaneously diagonalized by the similarity transformation associated to the matrix H (we will prove that H is orthogonal). We analyze the structure of $\mathcal{H}_n$, and we prove that this class includes the set of all the real symmetric circulant matrices. This fact allows us to determine new effective preconditioners of Toeplitz systems, which, having a higher number of free parameters, may further reduce the condition number of the preconditioned matrix obtained by using circulant matrices. Some numerical experiments reported at the end of §3 confirm this property. Actually, we extend to the class $\mathcal{H}_n$ the results of [15], [22], and [12], by proving that the eigenvalues of the preconditioned matrix are clustered around 1, and obtain algorithms having a better convergence, keeping the same computational cost per step. In fact, as it is easily verifiable (compare also, [11] and [23]), the discrete Hartley transform of real vector having n components can be computed with a cost of $\frac{3}{2}n\log n$ additions and $n\log n$ multiplications (where we assume that n is an integer power of 2, the cost bound is given up to additive constants, and all the logarithms are to the base 2). The above cost bound is the same cost bound of performing a discrete Fourier transform (DFT) of a real vector.

2. The algebra $\mathcal{H}_n$. Let us start by analyzing some properties of the matrices C, S, and H defined in §1.

LEMMA 1. *The following relations hold:*

$$\begin{aligned} &C^2+S^2=nI,\\ &CS=SC=0,\\ &HH^T=H^2=I,\\ &JS=SJ=-S,\\ &JC=CJ=C,\\ &HJ=JH, \end{aligned}$$

where J is the permutation matrix

$$J=\begin{pmatrix} 1 & 0 & \dots & 0 & 0\\ 0 & \iddots & \iddots & \iddots & 1\\ \vdots & \iddots & \iddots & \iddots & 0\\ 0 & \iddots & \iddots & \iddots & \vdots\\ 0 & 1 & 0 & \dots & 0 \end{pmatrix}.$$

Proof. From the relations $F^HF=FF^H=I$ and $F^2=J$ we have $C^2+S^2=nI$, $CS=SC=0$. Therefore $H^TH=\frac{1}{n}(C^2+S^2+CS+SC)=I$. From the properties of trigonometric functions, we obtain $SJ=JS=-S$, $JC=CJ=C$, whence $HJ=JH$. □

Let $\mathcal{A}_n$ and $\mathcal{B}_n$ denote the sets of real symmetric and skewsymmetric circulant $n\times n$ matrices, respectively; i.e.,

$$\begin{aligned} &\{\mathrm{Circ}(c_0,c_1,\dots,c_{n-1}),\ c_i\in\mathbf{R},\ i=0,\dots,n-1,\ c_i=c_{n-i},\ i=1,\dots,\lfloor n/2\rfloor\},\\ &\{\mathrm{Circ}(0,c_1,\dots,c_{n-1}),\ c_i\in\mathbf{R},\ i=1,\dots,n-1,\ c_i=-c_{n-i},\ i=1,\dots,\lfloor n/2\rfloor\}, \end{aligned}$$

where $\mathbf{R}$ is the real field, and define the following class of $n\times n$ matrices

$$\mathcal{H}_n=\{E=A+JB:\ A\in\mathcal{A}_n,\ B\in\mathcal{B}_n\}.$$

Any matrix $E\in\mathcal{H}_n$ can be expressed as the sum of two independent matrices, the first being a symmetric circulant, the second a special Hankel matrix (a Hankel matrix has entries depending only on the sum of their subscripts), for instance, for $n=5$,

$$E=\begin{pmatrix} a_0 & a_1 & a_2 & a_2 & a_1\\ a_1 & a_0 & a_1 & a_2 & a_2\\ a_2 & a_1 & a_0 & a_1 & a_2\\ a_2 & a_2 & a_1 & a_0 & a_1\\ a_1 & a_2 & a_2 & a_1 & a_0 \end{pmatrix}+\begin{pmatrix} 0 & b_1 & b_2 & -b_2 & -b_1\\ b_1 & b_2 & -b_2 & -b_1 & 0\\ b_2 & -b_2 & -b_1 & 0 & b_1\\ -b_2 & -b_1 & 0 & b_1 & b_2\\ -b_1 & 0 & b_1 & b_2 & -b_2 \end{pmatrix}.$$

Now we can prove the following results.

THEOREM 1. *For any matrix $E \in \mathcal{H}_n$ the following relations hold:*

$$HEH = \mathrm{diag}(d_0, \ldots, d_{n-1}), \qquad (d_i) = \sqrt{n} HE\mathbf{e}_1.$$

Proof. Let A be any circulant matrix having real entries so that, by (1.1), $F^H AF = D_A$ where D_A is a diagonal matrix; moreover,

$$F^H AF = \frac{1}{n}(SAS + CAC) + \mathrm{i}\frac{1}{n}(CAS - SAC),$$

whence

$$\mathrm{Re}(D_A) = \frac{1}{n}(SAS + CAC) \quad \text{and} \quad \mathrm{Im}(D_A) = \frac{1}{n}(CAS - SAC).$$

On the other hand,

$$HAH = \frac{1}{n}(C+S)A(C+S) = \frac{1}{n}(CAC + SAS + CAS + SAC),$$

and applying Lemma 1 yields $CAS + SAC = J(CAS - SAC)$, whence

$$HAH = \mathrm{Re}(D_A) + J\mathrm{Im}(D_A).$$

For a general matrix $E \in \mathcal{H}_n$, $E = A + JB$, $A \in \mathcal{A}_n$, $B \in \mathcal{B}_n$, since F is a unitary matrix and A and B are real symmetric and real skewsymmetric matrices, respectively, we have that the diagonal matrices $D_A = F^H AF$ and $D_B = F^H BF$ have real and imaginary entries, respectively, i.e., $\mathrm{Im}(D_A) = 0$, $\mathrm{Re}(D_B) = 0$. This fact, together with Lemma 1, implies that

$$HEH = HAH + HJBH = HAH + JHBH = \mathrm{Re}(D_A) + \mathrm{Im}(D_B) = D.$$

The equation $(d_i) = \sqrt{n}HE\mathbf{e}_1$ is obtained by applying both members of the equation $HEH = \mathrm{diag}(d_0, \ldots, d_{n-1})$ to the vector $\mathbf{e} = (1, 1, \ldots, 1)^T$ since $H\mathbf{e} = \sqrt{n}\mathbf{e}_1$. □

Observe that, since $\mathcal{H}_n = \mathcal{A}_n \oplus J\mathcal{B}_n$ (where $\oplus$ denotes the direct sum of subspace and $J\mathcal{B}_n$ is the subspace made up by the matrices JB, $B \in \mathcal{B}_n$), we have that $\dim \mathcal{H}_n = \dim \mathcal{A}_n + \dim \mathcal{B}_n = n$. In other words, the n-dimensional algebra $\mathcal{H}_n$ is isomorphic to the algebra of $n \times n$ diagonal matrices over the real field.

The above theorem allows us to devise methods to solve any system $E\mathbf{x} = \mathbf{b}$, $E \in \mathcal{H}_n$, in $O(n \log n)$ operations. This is due to the fact that for real vectors, the DHT can be computed with the same number of arithmetic operations in the real field, used in the computation of the DFT by means of fast Fourier transform (FFT) algorithms. In fact, either the DFT or the DHT of a real vector of n components can be computed in at most $\frac{3}{2}n \log_2 n$ additions and $n \log_2 n$ multiplications of real numbers, where n is an integer power of 2 [11], [23].

3. An application to preconditioning Toeplitz systems. Consider a linear system $T_n\mathbf{x} = \mathbf{b}$, where $T_n = (t_{i,j})$ is a real symmetric $n \times n$ Toeplitz matrix, that is, $t_{i,j} = t_{|i-j|}$, $i, j = 0, \ldots, n-1$. Toeplitz matrices arise in many applications (see [10]) and concretely fast algorithms for the solution of these systems are extremely important.

The conjugate gradient method [18] is particularly suitable for Toeplitz systems. In fact, the most expensive stages at each step of this algorithm consists in the computation of a matrix-by-vector product, where the matrix is Toeplitz. Due to this structure, such a product can be computed in $O(n \log n)$ operations. However, the algorithm takes n steps to arrive to completion so that the overall cost would be too high. A preconditioning strategy can be used for obtaining a good approximation to the solution in a constant number of steps. That is, a real symmetric positive-definite matrix S_n (the *preconditioner*) should be determined so that

(a) S_n is easily invertible (possibly in $O(n \log n)$ arithmetic operations);

(b) The matrix S_n is close to T_n, that is, the eigenvalues of $S_n^{-1}T_n$ are clustered around 1.

In this case, the preconditioned conjugate gradient method computes an approximation to the solution in a number of steps that depends on the number of eigenvalues of $S_n^{-1}T_n$ clustered around 1. The cost of each step is increased by the cost of solving a linear system with the matrix S_n.

Condition (a) is easily satisfied if we choose S_n in a matrix algebra made up by matrices which are simultaneously diagonalizable by a similarity transformation associated to a fast transform, for instance DFT, sine transform, or DHT. Condition (b) has been proved for S_n belonging to the classes $\mathcal{C}_n$ and $\mathcal{T}_n$ [15], [12], [6]. In this section we prove that condition (b) is satisfied by the matrix that solves the following minimum problem:

$$\min_{S_n \in \mathcal{S}_n} ||S_n - T_n||_F, \tag{3.1}$$

where $||X||_F = \sqrt{\sum_{i,j} |x_{i,j}|^2}$ denotes the Frobenius norm. In [12] and [6] problem (3.1) has been treated in the cases where $\mathcal{S}_n = \mathcal{C}_n$ and $\mathcal{S}_n = \mathcal{T}_n$, respectively. In this section we deal with problem (3.1) in the case where $\mathcal{S}_n = \mathcal{H}_n$. Moreover, since real symmetric circulant matrices are a subalgebra of $\mathcal{H}_n$, choosing S_n in the class $\mathcal{H}_n$ allows us to obtain a better approximation in the sense of (3.1).

THEOREM 2. *The matrix $S_n = A_n + JB_n \in \mathcal{H}_n$ defined by*

$$a_0 = t_0,$$
$$a_i = \frac{(n-i)t_i + it_{n-i}}{n}, \quad b_i = \frac{t_i - t_{n-i}}{n}, \quad i = 1, \ldots, n-1$$

minimizes $||S_n - T_n||_F$ in the class $\mathcal{H}_n$, where $T_n = (t_{i,j})$, $t_{i,j} = t_{|i-j|}$.

Proof. The result is obtained by setting to zero the partial derivatives of $||S_n - T_n||_F$ with respect to a_0, a_i, b_i, $i = 1, \ldots, n-1$. □

Note that the values b_i, $i = 1, \ldots, n-1$, can be seen as correction terms to the preconditioner of [12] when the matrix T_n is not circulant.

In the following, under the same assumptions on the matrix T_n of [12] and [6], we study the spectral properties of the matrix $S_n^{-1}T_n$. More precisely, we suppose that the matrices T_n, $n = 1, 2, \ldots$, are finite sections of a single, infinite, symmetric matrix T_∞, generated by the real valued function $f(z) = \sum_{j=-\infty}^{+\infty} t_j z^j$ defined on the unit circle in the complex plane and that f belongs to the *Wiener class*, that is, $\sum_{j=-\infty}^{+\infty} |t_j| < +\infty$. The following result is an extension to the class $\mathcal{H}_n$ of analogous results holding for $\mathcal{C}_n$ and $\mathcal{T}_n$.

THEOREM 3. *Let T_n be a real symmetric Toeplitz matrix defined, as a finite section, from the function $f(z) = \sum_{j=-\infty}^{+\infty} t_j z^j$ belonging to the Wiener class. Then*

(i) *for any $\epsilon > 0$ there exists an integer n_0 such that for any $n > n_0$ the spectrum of S_n lies in the interval $[f_{\min} - \epsilon, f_{\max} + \epsilon]$, where $f_{\min}$ and $f_{\max}$ are the extremal values of f on the unit circle;*

(ii) *the spectrum of $S_n^{-1}T_n$ is clustered around 1, that is, for any $\epsilon > 0$ there exist integers k and n_0 such that for any $n > n_0$ the number of eigenvalues λ_i of $S_n^{-1}T_n$ such that $|\lambda_i - 1| > \epsilon$ is less than k.*

Proof. Concerning (i), we observe that if $\lambda_j(S_n)$ denotes the jth eigenvalue of S_n, then

$$\begin{aligned}\lambda_j(S_n) &= t_0 + \frac{2}{n}\sum_{k=1}^{n-1} t_k \left((n-k)\cos\frac{2\pi k(j-1)}{n} - \sin\frac{2\pi k(j-1)}{n}\right)\\ &= \mathrm{Re}\left(\sum_{k=-n+1}^{n-1} t_k e^{\mathrm{i}\frac{2\pi k(j-1)}{n}}\right) - \frac{2}{n}\sum_{k=1}^{n-1} t_k k\cos\frac{2\pi k(j-1)}{n}\\ &\quad - \frac{2}{n}\sum_{k=1}^{n-1} t_k \sin\frac{2\pi k(j-1)}{n}.\end{aligned}$$

Property (i) follows since the sequences

$$\left\{\frac{2}{n}\sum_{k=1}^{n-1} t_k k\cos\frac{2\pi k(j-1)}{n}\right\} \quad \text{and} \quad \left\{\frac{2}{n}\sum_{k=1}^{n-1} t_k \sin\frac{2\pi k(j-1)}{n}\right\}$$

tend to zero. Concerning (ii), it is sufficient to show that the eigenvalues of $S_n^{-1}(T_n - S_n)$ are clustered around zero, in fact, $S_n^{-1}T_n = S_n^{-1}(T_n - S_n) + I$. Let $\epsilon > 0$ be fixed. Since f belongs to the Wiener class, we can choose N such that $\sum_{j=N+1}^{+\infty} |t_j| < \epsilon$. The matrix $T_n - S_n = T_n - A_n - JB_n$, can be split as

$$W_n^{(N)} + Z_n^{(N)} + E_n^{(N)},$$

where the first two matrices agree with the $(n-N) \times (n-N)$ leading principal submatrices of $T_n - A_n$ and JB_n, respectively. We observe that:

$$\begin{aligned}&\mathrm{rank}(E_n^{(N)}) \le 2N,\\ &||W_n^{(N)}||_1 \le \frac{2}{n}\sum_{k=1}^{n-N-1} k|t_{n-k} - t_k| \le \frac{2}{n}\sum_{k=1}^{N} k|t_k| + 4\sum_{k=N+1}^{n} |t_k|,\\ &||Z_n^{(N)}||_1 \le \frac{2}{n}\sum_{k=1}^{\lfloor n-1/2\rfloor} |t_k - t_{n-k}| \le \frac{2}{n}\sum_{k=1}^{n-1} |t_k|,\end{aligned}$$

where $||X||_1 = \max_j \sum_i |x_{i,j}|$ denotes the 1-norm. Now, let $n_0 > N$ be such that $\frac{1}{n_0}\sum_{k=1}^{N} k|t_k| < \epsilon$ and $\frac{1}{n_0}\sum_{k=1}^{n_0-1} |t_k| < \epsilon$. Then, for any $n > n_0$ we have

$$||W_n^{(N)} + Z_n^{(N)}||_1 \le ||W_n^{(N)}||_1 + ||Z_n^{(N)}||_1 \le 8\epsilon.$$

Hence, by Cauchy interlace theorem [24], the eigenvalues of $T_n - S_n$ are clustered around zero, except at most $k = 2N$ of them. Applying the Courant–Fischer minimax characterization [24] to the matrix $S_n^{-1}(T_n - S_n)$, we have

$$\lambda_i(S_n^{-1}(T_n - S_n)) < \frac{\lambda_i(T_n - S_n)}{f_{\min}},$$

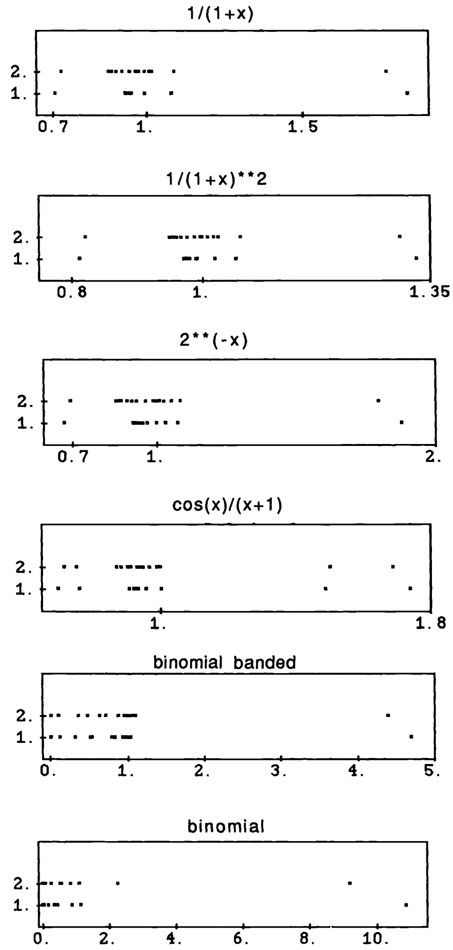


FIG. 1.

for n sufficiently large, hence, even the spectrum of $S_n^{-1}(T_n - S_n)$ is clustered around zero, that is, for any $\epsilon > 0$ there exist integers k and n_0 such that for any $n > n_0$ the number of eigenvalues λ_i of $S_n^{-1}T_n$ which verify $|\lambda_i - 1| > \epsilon$ is less than k. □

The results of some numerical experiments for different Toeplitz matrices confirm the theoretical results. In Fig. 1 we report the eigenvalues location of the preconditioned matrices obtained by applying the circulant preconditioner C of [12] (denoted by 1) and the Hartley preconditioner H (denoted by 2) to the matrices whose (i,j)-entry is $t_{i,j} = f(|j-i|)$. We consider the same size $n = 15$ and the same functions of [12]: $f(x) = (1/1+x), f(x) = (1/(1+x)^2), f(x) = 2^{-x}, f(x) = (\cos x/1+x)$. We also applied the preconditioner to the very ill conditioned matrix defined by the function

$$f(x) = \begin{cases} \binom{2k}{k-x} & \text{if } x \leq k, \\ 0 & \text{if } x > k, \end{cases}$$

in the cases where $k = 2$ and $k = 16$, denoted as banded binomial and binomial functions. In all the experiments the spectrum of $H^{-1}T$ completely lies within that of $C^{-1}T$, and hence the condition number of $H^{-1}T$ is smaller.

REFERENCES

[1] G. AMMAR AND P. GADER, *A variant of the Gohberg–Semencul formula involving circulant matrices*, SIAM J. Matrix Anal. Appl., 12 (1992), pp. 534–540.

[2] J. L. BARLOW, *Error analysis of update methods for the symmetric eigenvalues problem*, Tech. Rep. CS-91-09, Dept. of Computer Science, Pennsylania State Univ., University Park, PA.

[3] D. BINI, *On a class of matrices related to Toeplitz matrices*, Tech. Rep. 83-5, Dept. of Computer Science, State University of New York at Albany, 1983.

[4] D. BINI AND M. CAPOVANI, *Spectral and computational properties of band symmetric Toeplitz matrices*, Linear Algebra Appl., 52 (1983), pp. 99–126.

[5] ———, *Tensor rank and border rank of band Toeplitz matrices*, SIAM J. Comput., 16 (1987), pp. 252–258.

[6] D. BINI AND F. DI BENEDETTO, *A new preconditioner for the parallel solution of positive definite Toeplitz systems*, in second Ann. Symp. Parallel Algorithms and Architecture, Crete, Greece, 1990, pp. 220–223.

[7] D. BINI AND V. PAN, *Numeric and Algebraic Computations with Matrices and Polynomials, Vol.* 1: *Fundamental Algorithms*, Birkhäuser, Boston, to appear.

[8] R. BRACEWELL, *Discrete Hartley transform*, J. Opt. Soc. Amer., 73 (1983), pp. 1832–1835.

[9] ———, *The fast Hartley transform*, Proc. IEEE, 72 (1984), pp. 1010–1018.

[10] J. BUNCH, *Stability of methods for solving Toeplitz systems of equation*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.

[11] O. BUNEMAN, *Conversion of* FFT's *to fast Hartley transform*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 624–638.

[12] T. CHAN, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

[13] R. CHAN, *The spectrum of a family of circulant preconditioned Toeplitz systems*, SIAM J. Numer. Anal., 26 (1989), pp. 503–506.

[14] ———, *Circulant preconditioners for Hermitian Toeplitz systems*, SIAM J. Matrix Anal. Appl., 4 (1989), pp. 542–550.

[15] R. CHAN AND G. STRANG, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 104–119.

[16] P. DAVIS, *Circulant Matrices*, John Wiley, New York, 1979.

[17] J. GRCAR AND A. SAMEH, *On certain parallel Toeplitz linear system solvers*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 238–256.

[18] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, The John Hopkins University Press, Baltimore, Maryland, 1983.

[19] T. HUCKLE AND A. HUBLAND, *Circulant and skewcirculant matrices for solving Toeplitz matrix problems*, in Proc. Copper Mountain Conf. Iterative Methods, Colorado, 1990.

[20] T. KU AND C.-C. J. KUO, *Design and Analysis of Toeplitz Preconditioners*, in Proc. ICASSP-90, Albuquerque, New Mexico 1990.

[21] W. PRESS, B. FLANNERY, S. TENKOLSKY, AND W. VETTERLING, *Numerical Recipes—the Art of Scientific Computing*, Cambridge University Press, London, 1986.

[22] G. STRANG, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.

[23] H. SORENSEN, D. JONES, C. BURRUS, AND M. HEIDEMAN, *On computing the discrete Hartley transform*, IEEE Trans. Acoust., Speech and Signal Processing, ASSP-33 (1985), pp. 1231–1238.

[24] J. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 508–520, April 1993

ROBUST STABILITY AND DIAGONAL LIAPUNOV FUNCTIONS*

EUGENIUS KASZKUREWICZ† AND AMIT BHAYA†

Abstract. This paper shows that diagonal Liapunov functions play an essential role in the robust stability problem of linear autonomous continuous and discrete-time systems in state-space form subjected to a broad class of perturbations. The results obtained generalize a well-known result of Persidskii and are related to a variety of similar results in the literature. Techniques are suggested for applying the results to the problem of evaluating robust stability bounds in different contexts. Several examples are given.

Key words. nonlinear systems, stability criteria, robustness, Liapunov methods, diagonal stability

AMS(MOS) subject classifications. 93D09, 93D05, 93D20

1. Introduction.

There is an extensive literature concerning the study of robust stability of dynamical systems. Many different approaches have been used; see, for example, [4], [8], and [36], and the references therein. The present paper is concerned with conditions under which the asymptotically stable linear systems in state-space form

$$\dot{x}(t) = Ax(t), \qquad t \geq 0, \tag{1}$$

$$x(k+1) = Ax(k), \qquad k = 0, 1, 2, \ldots, \tag{2}$$

with $x(t), x(k) \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$, remain globally asymptotically stable when subjected to some classes of nonlinear and time-varying perturbations, namely, *state-dependent* and *parametric* perturbations.

Liapunov's direct method is used to show that for the classes of perturbations under consideration, diagonal Liapunov functions play an essential role in the robustness problem. In other words, if the nominal system matrices $A = (a_{ij})$ or $|A| = (|a_{ij}|)$ are *diagonally stable* or, equivalently, if the Liapunov equation admits a positive diagonal solution, then the robustness of the corresponding linear systems (1) and (2) is ensured by these diagonal Liapunov functions.

Since the results presented in this paper for different classes of systems are all based on the existence of a diagonal solution to Liapunov's equation, they can be viewed as extensions of the result on absolute stability given by Persidskii [33].

Our robustness results are easy to state since they are based on the diagonal stability of certain test matrices derived from the nominal systems (1) and (2). In some cases the test matrices are nonnegative or are M-matrices, and in these cases diagonal stability is equivalent to Schur or Hurwitz stability, so that we can use some of the known criteria for linear systems in our nonlinear robust stability problem (see §3). This is useful, especially in the problem of the evaluation of robustness bounds.

2. Stability conditions for the perturbed systems.

In this section we consider the nominal linear systems in state-space form $\dot{x}(t) = Ax(t)$ and $x(k+1) = Ax(k)$, subject to two types of nonlinear and time-varying perturbations: *state-dependent* perturbations [12] and *parameter variations* in matrices $A = (a_{ij})$. The

* Received by the editors December 27, 1989; accepted for publication (in revised form) August 29, 1991.

† Department of Electrical Engineering, Federal University of Rio de Janeiro, COPPE/UFRJ, P.O. Box 68504, 21945970 Rio de Janeiro, Brazil (coe10027@ufrj.bitnet). This research was partially supported by CNPq.

effect of these two types of perturbations is described by nonlinear models, as discussed below, and the robust stability conditions are given for each class of perturbations.

2.1. State-dependent perturbations. Let the nominal systems (1) and (2) be subjected to state-dependent perturbations. We assume that the perturbed systems are described (respectively) by the equations below.

$$\Sigma_c : \dot{x}_i(t) = \sum_{j=1}^{n} a_{ij} f_j(x_j(t)), \qquad i = 1, \ldots, n, \tag{3}$$

$$\Sigma_d : x_i(k+1) = \sum_{j=1}^{n} a_{ij} \phi_j(x_j(k)), \qquad i = 1, \ldots, n. \tag{4}$$

The continuous functions $f_i(\cdot) : \mathbb{R} \to \mathbb{R}$ and $\phi_i(\cdot) : \mathbb{R} \to \mathbb{R}$ describe the perturbations on the state variables $x_i(t)$ and $x_i(k)$, respectively. In addition, in our analysis we assume that the classes of functions are such that

$$f_i(\cdot) \in S_c := \{f(\cdot) | f : \mathbb{R} \to \mathbb{R}; f(\xi) \cdot \xi > 0; f(0) = 0\}, \tag{5}$$

$$\phi_i(\cdot) \in S_d := \{\phi(\cdot) | \phi : \mathbb{R} \to \mathbb{R}; |\phi(\xi)| \leq |\xi|; \phi(0) = 0\}, \tag{6}$$

with the additional condition

$$\forall i, \quad \int_0^{x_i} f_i(\tau) \to \infty \quad \text{as } |x_i| \to \infty \tag{7}$$

and pertinent smoothness conditions on the functions $f_i(\cdot)$ to assure the existence and uniqueness of a solution for (3). The class of perturbed systems described by (3) (respectively, (4)) for $f_i(\cdot) \in S_c$, for all i (respectively, $\phi_i(\cdot) \in S_d$, for all i) is denoted by Σ_c (respectively, Σ_d).

Note that the state-dependent perturbations on the continuous-time model are of the infinite-sector type (one sided and unlimited). In the discrete-time case it is natural to consider the (normalized) $[-1, 1]$ sector (class S_d). Note also that $x(t) = 0$, for all t (respectively, $x(k) = 0$, for all k) are solutions of (3) (respectively (4)), and throughout this paper we will be concerned with the (asymptotic) stability of this zero solution, also called an equilibrium.

In order to state the results, we define two classes of constant square $n \times n$ matrices:

$$\mathcal{D}_c := \{A = (a_{ij}) : \exists \text{ diagonal} P > 0 \text{ s.t. } PA + A^T P = -Q < 0 \text{ for some } Q > 0\}.$$
$$\mathcal{D}_d := \{A = (a_{ij}) : \exists \text{ diagonal} P > 0 \text{ s.t. } A^T PA - P = -Q < 0 \text{ for some } Q > 0\}.$$

In other words, $\mathcal{D}_c$ and $\mathcal{D}_d$ define the classes of matrices that satisfy Liapunov's equations with a positive diagonal matrix P; these matrices are called *diagonally stable* [3].

We will also refer to the sets of matrices that satisfy Liapunov's equations with some positive definite matrices P and Q (with P not necessarily diagonal) by $\mathcal{H}$ and $\mathcal{S}$, i.e., the sets of Hurwitz and Schur matrices (i.e., stable matrices).

For the class Σ_c of perturbed systems (3), obtained by nonlinear state-dependent perturbations (belonging to class S_c) of the nominal system, Persidskii [33] proved, roughly speaking, that a diagonally stable nominal system is robustly stable when subject to a class of nonlinear perturbations in the state vector. More precisely, we have the following theorem.

THEOREM 2.1. *The equilibrium $x = 0$ of all systems in the class Σ_c is globally asymptotically stable if the nominal system matrix $A = (a_{ij})$ belongs to the class $\mathcal{D}_c$, i.e., if the nominal system is diagonally stable.*

Proof. Since $A \in \mathcal{D}_c$, there exists a positive diagonal matrix D such that $A^TD + DA$ is negative definite. Let $D = \operatorname{diag}(p_1/2, \dots, p_n/2)$, and let

$$V(x) = \sum_{i=1}^{n} p_i \int_0^{x_i} f_i(\tau)d\tau \tag{8}$$

be the diagonal Liapunov function candidate. Computing $\dot{V}$ along the trajectories of (3) gives

$$\dot{V}(x(t)) = f(x(t))^T(DA + A^TD)f(x(t)),$$

where $f(x)^T = (f_1(x_1), \dots, f_n(x_n))$. Since $A \in \mathcal{D}_c$, it follows that $\dot{V}(x(t)) < 0$ (is negative definite). □

Remark 2.1. A limitation of the nonlinearity-dependent Liapunov function (8) introduced by Persidskii is that it works only for the diagonal state-dependent nonlinearity $f(\cdot)$, so that if we admit only this type of Liapunov function, then we are restricted to considering models that do not permit independent (time-varying nonlinear) perturbations of components of the state vector and of the elements of the nominal system matrix A, unless rather restrictive conditions are imposed on the class of admissible perturbations. In §2.2 we introduce a more general class of perturbations that can accommodate these types of perturbations. It turns out, in fact, that it is enough to consider independent perturbations of the components of the state vector because this allows us to model perturbations in the entries of the nominal matrix as well.

Remark 2.2. The above result is also valid for the model given by the equations

$$\Sigma_c^* : \dot{x}_i(t) = f_i\left(\sum_{j=1}^{n} a_{ij}x_j(t)\right), \qquad i = 1, \dots, n. \tag{9}$$

This can be shown by using the change of variables $y(t) = Ax(t)$; then (9) can be rewritten as $\dot{y}_i(t) = \sum_{j=1}^{n} a_{ij}f_j(y_j(t))$, which is of the form (3). Note that the condition $\det A \neq 0$, required to define y as a change of variables, is implied by the assumption that the nominal system $\dot{x} = Ax$ is stable.

For the discrete-time case, the result equivalent to Theorem 2.1 was stated in [20] and is generalized below for time-varying nonlinearities satisfying

$$\forall i, \quad |\phi_i(x_i(k), k)| \leq |x_i(k)|.$$

In other words, the system considered is given by the equations

$$\overline{\Sigma}_d(k) : x_i(k+1) = \sum_{j=1}^{n} a_{ij}\phi_j(x_j(k), k), \qquad i = 1, \dots, n. \tag{10}$$

Hereafter, we will also denote the classes of functions $f_i(\cdot,\cdot)$ and $\phi_i(\cdot,\cdot)$ by $S_c(t)$ and $S_d(k)$:

$$S_c(t) := \{f(\cdot,\cdot) | f : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}; f(\xi, t)\xi > 0 \; f(0,t) = 0, \forall t\},$$
$$S_d(k) := \{\phi(\cdot,\cdot) | \phi : \mathbb{R} \times \mathbf{I} \to \mathbb{R}; |\phi(\xi, k)| \leq |\xi|; \phi(0,k) = 0, \forall k\}.$$

Remark 2.3. Note that a function $\phi(\xi, k) \in S_d(k)$ can be zero for several values of ξ and k (indeed, $\phi(\xi, k) \equiv 0$, for all ξ, for all k is in $S_d(k)$), in sharp contrast to an $f \in S_c(t)$.

Notation. Let the mapping Φ denote $(\phi_1, \ldots, \phi_n)^T$, where $\phi_i \in S_d(k)$, for all i. Thus we can write the class of perturbed systems $\overline{\Sigma}_d(k)$ in (10) as

$$x(k+1) = A\Phi(x(k), k), \qquad \Phi \in S_d(k)^n. \tag{11}$$

Also, in Lemma 2.2 below, we will use the slight generalization of the class $S_d(k)$ defined below:

(12)

$$S_d^g(k)^n := \{\Phi(\cdot, \cdot) | \phi_i : \mathbb{R}^n \times \mathbf{I} \to \mathbb{R}; |\phi_i(x, k)| \le |x_i|; \forall i, \forall x \in \mathbb{R}^n;\ \phi_i(0, k) = 0, \forall k\}.$$

LEMMA 2.2. *Let $\Phi \in S_d^g(k)^n$ and $P \in \mathbb{R}^{n \times n}$, P diagonal. Then*

$$\forall x \in \mathbb{R}^n, \qquad \Phi(x(k), k)^T P \Phi(x(k), k) \le x^T P x. \tag{13}$$

Proof.

$$\begin{aligned} \forall P \text{ diagonal}, \quad \forall \Phi \in S_d^g(k)^n, \quad \Phi(x(k), k)^T P \Phi(x(k), k) &= \sum_{i=1}^n p_i \phi_i(x(k), k)^2 \\ &= \sum_{i=1}^n p_i |\phi_i(x(k), k)|^2 \\ &\le \sum_{i=1}^n p_i x_i^2 \\ &= x^T P x. \qquad \square \end{aligned}$$

COROLLARY 2.3. *For all $A \in \mathbb{R}^{n \times n}$, $\Phi(Ax(k), k)^T P \Phi(Ax(k), k) \le x^T A^T P A x$.*

Proof. Replace x by Ax in Lemma 2.2. □

We can now state a discrete-time version of Theorem 2.1.

THEOREM 2.4. *The equilibrium $x = 0$ of all systems in the class $\overline{\Sigma}_d(k)$ is globally asymptotically stable if the nominal matrix $A = (a_{ij})$ belongs to class $\mathcal{D}_d$.*

Proof. Since $A \in \mathcal{D}_d$, there exists a positive diagonal matrix P such that $A^T PA - P$ is negative definite. Define the candidate Liapunov function $V(x) := x^T P x$, and let $\Delta V(x(k), k) := V(x(k+1)) - V(x(k))$. Then

$$\Delta V(x(k), k) = \Phi^T(x(k), k) A^T P A \Phi(x(k), k) - x^T(k) P x(k).$$

There are two cases to consider.

Case 1. $x(k) \neq 0$ and k such that $\Phi(x(k), k) = 0$. Then $\Delta V(x(k), k) = -x^T(k) P x(k) < 0$.

Case 2. $x(k)$ and k such that $\Phi(x(k), k) \neq 0$. In this case, we have

$$\begin{aligned} \Delta V(x(k), k) &= \Phi^T(x(k), k) A^T P A \Phi(x(k), k) - x^T(k) P x(k) \\ &\le \Phi^T(x(k), k) A^T P A \Phi(x(k), k) - \Phi^T(x(k), k) P \Phi(x(k), k) \\ &= \Phi^T(x(k), k)(A^T P A - P)\Phi(x(k), k) \\ &< 0, \end{aligned}$$

where the first inequality follows from Lemma 2.2 and the second inequality from the fact that $A \in \mathcal{D}_d$. □

Remark 2.4. Notice that Theorem 2.4 could have been stated for the class of *nondiagonal* nonlinearities $S_d^g(k)^n$ defined in (12) without any change in the proof. Furthermore, notice that the Liapunov function $V(x)$ defined in Theorem 2.4 is time invariant and works for the entire class $\overline{\Sigma}_d(k)$. This is discussed further in Remark 2.7 and at the end of this section.

Remark 2.5. By Corollary 2.3, Theorem 2.4 also applies to the model

$$x_i(k+1) = \phi_i\left(\sum_{j=1}^{n} a_{ij}x_j(k), k\right), \qquad i = 1, \ldots, n. \tag{14}$$

This result also appears in [28] (also see §4, Example 3, below).

2.2. Parameter and state-dependent perturbations. As was pointed out in Remark 2.1, using the classical Persidskii model (3) and the associated Liapunov function (8) makes it difficult to accommodate general classes of parameter perturbations and time-varying nonlinearities that act simultaneously on the nominal linear model. We now consider more general models that can accommodate these types of perturbations; they are given by

$$\Sigma_c(t) : \dot{x}_i(t) = \sum_{j=1}^{n} a_{ij}f_{ij}(x_j(t), t), \qquad i = 1, \ldots, n, \tag{15}$$

$$\Sigma_d(k) : x_i(k+1) = \sum_{j=1}^{n} a_{ij}\phi_{ij}(x_j(k), k), \qquad i = 1, \ldots, n, \tag{16}$$

where the functions $f_{ij}(\cdot,\cdot)$ and $\phi_{ij}(\cdot,\cdot)$ belong to the classes $S_c(t)$ and $S_d(k)$, respectively. The classes of perturbed systems described by (15) and (16) for perturbations in the classes $S_c(t)$ and $S_d(k)$ are denoted by $\Sigma_c(t)$ and $\Sigma_d(k)$, respectively.

Remark 2.6. Notice that parameter- and time-dependent perturbations on the elements of matrix A of the type $(a_{ij} + \delta a_{ij}(t))$ (additive) and $(a_{ij} \times \delta a_{ij}(t))$ (multiplicative), first considered in the linear case in [31], [37], are special cases of models (15) and (16), since

$$\begin{aligned} &(a_{ij} + \delta a_{ij}(t))x_j = a_{ij}[x_j + (\delta a_{ij}(t)/a_{ij})x_j] = a_{ij}f_{ij}(x_j, t), \qquad a_{ij} \neq 0,\\ &(a_{ij} \times \delta a_{ij}(t))x_j = a_{ij}(\delta a_{ij}(t) \times x_j) = a_{ij}f_{ij}(x_j, t). \end{aligned}$$

However, as will be seen below, the main thrust of the theorems that follow is to provide simple conditions on test matrices derived in a straightforward manner from the nominal system matrices to ensure robust stability to the wide classes of nonlinear time-varying perturbations $S_c(t)$ and $S_d(k)$. Of course, perturbations that are less structured than these can also be expressed by models (15) and (16), so that Theorems 2.5 and 2.6 below are quite general. In order to state and prove these theorems we define matrices $\overline{A}$ and $|A|$ as

$$\begin{aligned} &\overline{A} = (\overline{a}_{ij}), \qquad \overline{a}_{ij} = \begin{cases} a_{ij}, & i = j,\\ |a_{ij}|, & i \neq j, \end{cases}\\ &|A| = (|a_{ij}|). \end{aligned} \tag{17}$$

THEOREM 2.5. *If* (i) *the nominal system matrix A is such that* $(a_{ii} < 0)$, $i = 1, \dots, n$ *and $\overline{A}$ belongs to class $\mathcal{D}_c$, and* (ii) *there exist continuous functions $f_i(\cdot) \in S_c$ that satisfy* (7) *and such that the following elementwise diagonal-dominance-type conditions hold for all j,*

$$|f_{ij}(\xi,t)| \le |f_j(\xi)| \le |f_{jj}(\xi,t)|, \quad \forall i \text{ and } \forall t \ge 0, \tag{18}$$

then the equilibrium $x = 0$ of all systems in the class $\Sigma_c(t)$ is globally asymptotically stable.

Proof. Using

$$V(x) = \sum_{i=1}^{n} p_i \int_0^{x_i} f_i(\tau)\mathrm{d}\tau, \quad p_i > 0, \quad i = 1, \dots, n \tag{19}$$

and computing $\dot{V}$ along the trajectories of (15) gives

$$\dot{V}(x,t) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i a_{ij} f_i(x_i) f_{ij}(x_j,t).$$

From condition (18)

$$\dot{V}(x,t) \le \sum_{i=1}^{n}\sum_{j=1}^{n} p_i \bar{a}_{ij} |f_i(x_i)|.|f_j(x_j)|. \tag{20}$$

Let $x = (x_1, \dots, x_n)^{\mathrm{T}}$, and let the vectorial norm $[\![\cdot]\!]$ be defined as

$$[\![x]\!]^{\mathrm{T}} := (|x_1|, \dots, |x_n|). \tag{21}$$

Then inequality (20) can be rewritten in the form

$$\dot{V}(x,t) \le [\![f(x)]\!]^{\mathrm{T}} \, (\overline{A}^{\mathrm{T}} D + D\overline{A})[\![f(x)]\!],$$

and since $\overline{A} \in \mathcal{D}_c$ (with $D = \mathrm{diag}(p_i/2)$), $\dot{V}(x,t) < 0$. □

Remark 2.7. Note that in (18) it is required only that the functions $f_{ii}(\cdot,\cdot)$ belong to class $S_c(t)$. Furthermore, although the Liapunov function in (19) looks very similar to the one in (8), there is a fundamental difference between the two. Equation (19) defines a *single time-invariant* Liapunov function (defined in terms of the *fixed* $f_i \in S_c(t)$) that *simultaneously* proves stability of the entire class of nonlinear time-varying perturbed systems $\Sigma_c(t)$. Equation (8), on the other hand, actually defines an entire class of Liapunov functions, each one tailored to prove the stability of a particular perturbation of the nominal system (namely, the one defined by the f_i that appear in the perturbed model). In other words, (19) defines a time-invariant and perturbation-invariant Liapunov function (and this is why it works to prove the stability of the time-varying perturbed systems), whereas (8) defines a Liapunov function that, although time invariant, is *not* perturbation invariant (i.e., not simultaneous) and hence would not work without additional hypotheses for time-varying perturbations. From this discussion, we see that the Liapunov function of Theorem 2.4 is also perturbation invariant and hence works for the time-varying case (see Remark 2.4).

Remark 2.8. In Theorem 2.5, given our hypothesis that $\overline{A} \in \mathcal{D}_c$, we must assume that the $a_{ii}, i = 1, \dots, n$ are negative, since this is a necessary condition for $\overline{A} \in \mathcal{D}_c$.

For the class of discrete-time systems $\Sigma_d(k)$ the corresponding result is the following theorem.

THEOREM 2.6. *The equilibrium $x = 0$ of all systems in the class $\Sigma_d(k)$ is globally asymptotically stable if the matrix $|A|$ belongs to class $\mathcal{D}_d$.*

Proof. If we use the diagonal Liapunov function $V(x) = \sum_{i=1}^n p_i x_i^2$ for the system (16), $\Delta V(x(k), k)$ has the form

$$\Delta V(x(k), k) = \sum_{i=1}^{n} p_i[x_i^2(k+1) - x_i^2(k)],$$

and, since $\phi_{ij}(\cdot,\cdot) \in S_d(k)$, one has

$$x_i^2(k+1) \leq \left(\sum_{j=1}^{n} |a_{ij}\phi_{ij}(x_j(k), k)|\right)^2 \leq \left(\sum_{j=1}^{n} |a_{ij}||x_j(k)|\right)^2.$$

Using the vectorial norm $[\![\cdot]\!]$ defined in (21), we obtain

$$\Delta V(x(k), k) \leq [\![x(k)]\!]^{\mathrm{T}} \left(|A^{\mathrm{T}}| \cdot P \cdot |A| - P\right)[\![x(k)]\!].$$

Since $|A| \in \mathcal{D}_d$ with $P = \mathrm{diag}(p_i)$, $\Delta V(x(k), k) < 0$. □

Remark 2.9. In Theorems 2.1–2.6, note that the corresponding Liapunov functions $V(x)$ satisfy $V(x) \to \infty$ with $\|x\| \to \infty$; this fact guarantees, as was stated in the theorems, global asymptotic stability in all cases. Furthermore, for Theorems 2.4–2.6, stated in the time-varying case, it is easy to see that V, $-\dot{V}$, and $-\Delta V$ are all positive and bounded away from zero, which is a requirement for asymptotic stability in the time-varying case.

Remark 2.10. Notice that, once again, the Liapunov function of the above theorem is *simultaneous* in the sense of Remark 2.7.

Remark 2.11. Theorem 2.6 also holds for the model

$$x_i(k+1) = \sum_{j=1}^{n} \phi_{ij}(a_{ij}x_j(k), k), \qquad i = 1, \ldots, n. \tag{22}$$

Remark 2.12. An interesting and related result holds for the following modification of model (12), in which different time-varying delays $(k - d_{ij}(k))$ are present between the state variables:

$$x_i(k+1) = \sum_{j=1}^{n} a_{ij}x_j(k - d_{ij}(k)), \tag{23}$$

where $0 \leq d_{ij}(k) \leq d$. Once again, diagonal stability of matrix $|A|$ with $P = \mathrm{diag}(p_i)$ assures the asymptotic stability of (23). In other words, given $|A| \in \mathcal{D}_d$, the linear system (2) has robust properties even in the presence of delays between the different state variables. This is a special case of a result due to Kaszkurewicz, Bhaya, and Šiljak [18], which uses a *diagonal* Liapunov function of the form $V(x) = \max_i\{p_i^{-1}|x_i|\}$.

Remark 2.13. Compare Theorem 2.1 with Theorem 2.5 and Theorem 2.4 with Theorem 2.6, and note that, since $(\overline{A} \in \mathcal{D}_c)$ implies $(A \in \mathcal{D}_c)$ and, similarly, $(|A| \in \mathcal{D}_d)$ implies $(A \in \mathcal{D}_d)$: under the hypotheses of Theorems 2.5 and 2.6, asymptotic stability in the presence of state-dependent perturbations is also guaranteed.

An interesting feature that emerges from the above analysis is that in Theorems 2.4, 2.5, and 2.6 *one Liapunov function* suffices to prove the asymptotic stability of an entire class of perturbed systems. However, Theorem 2.1 is different in that, although the condition $A \in \mathcal{D}_c$ guarantees asymptotic stability under state-dependent perturbations through the existence of a Liapunov function, the Liapunov function is constructed from these perturbations. This means that for a given nominal diagonally stable system, a Liapunov function that testifies to the stability for a given state-dependent perturbation does not do so for a different perturbation in the same sector. Thus by following [5] (see also [36]), we may label the time-invariant and perturbation-invariant Liapunov functions of Theorems 2.4, 2.5, and 2.6 as *simultaneous Liapunov functions.*

3. Discussion of the results. From Theorems 2.1 and 2.4–2.6 it becomes clear that *diagonal stability* of the matrices A, $\overline{A}$, and $|A|$ plays a crucial role in the (*simultaneous*) *robust stability* of the nominal systems $\dot{x}(t) = Ax(t)$ and $x(k+1) = Ax(k)$ subjected to perturbations in the classes defined above.

For an arbitrary matrix A, no simple characterization is known to check whether or not A is diagonally stable. Some numerical methods, however, are available [10], [16], [23], as are some characterizations for certain classes of matrices [2], [3], [6], [21], [29].

To apply Theorems 2.5 and 2.6, we can use several conditions equivalent to $\overline{A} \in \mathcal{D}_c$ and $|A| \in \mathcal{D}_d$, in particular, the following basic equivalences:

$$(\overline{A} \in \mathcal{D}_c) \Leftrightarrow (-\overline{A} \text{ is an M-matrix}) \Leftrightarrow (\overline{A} \in \mathcal{H}), \tag{24}$$

$$(|A| \in \mathcal{D}_d) \Leftrightarrow (I - |A| \text{ is an M-matrix}) \Leftrightarrow (|A| \in \mathcal{S}). \tag{25}$$

Suitable references for definitions and for more equivalences are [1], [18], and [35].

Notice that Theorems 2.5 and 2.6 are closely related to similar results obtained in the large-scale systems literature, which use properties of M-matrices. Because of the special structure of the models considered above and the particular choice of individual Liapunov functions for each coordinate i, the corresponding aggregate matrices ($\overline{A}$ and $|A|$) are *constructed directly* from the nominal systems (1) and (2).

In the general context of large-scale systems, the robust characteristics provided by M-matrices are known [35], and since the stability of M-matrices is equivalent to diagonal stability by (24) and (25), one can use this equivalence for the evaluation of robustness bounds by means of an interval matrix result for M-matrices [9], referenced in [25].

Robustness bounds can be derived on the basis of characterizations of diagonal stability of the matrices A, $\overline{A}$, and $|A|$. We give some instances of this general idea below.

In the case of Theorem 2.1, for example, the robustness bounds for the perturbations are given directly by the infinite sectors. This is equivalent to saying that if ($A \in \mathcal{D}_c$), then ($AK \in \mathcal{D}_c$) for any positive diagonal matrix $K = \operatorname{diag}(k_i)$, i.e., A is D-stable.

In the case of Theorem 2.4, robustness bounds on parameter and state-dependent perturbations are obtained by evaluating the set of positive diagonal matrices $K = \operatorname{diag}(k_i)$ for which ($AK \in \mathcal{D}_d$). This gives a region of admissible sectors for the nonlinearities $\phi_i(x_i(k), k)$ and for time-varying parameter perturbations as well.

In the case of Theorems 2.5 and 2.6, the robustness bounds are evaluated as in Theorem 2.4 and the evaluation of the region for the sectors to which the state-dependent perturbations and parametric time-dependent perturbations must belong is based on the diagonal stability of the matrices $(\overline{A} \circ K)$ and $(|A| \circ K)$, where $K = (k_{ij})$ is a matrix with positive entries and "$\circ$" denotes the elementwise (or Hadamard) product of matrices ($\overline{A}$ and K) and ($|A|$ and K), respectively.

If the nominal system matrix is not diagonally stable, we can, by means of state feedback, determine a new nominal system matrix $(A+BF)$ such that $(A+BF) \in \mathcal{D}_c$ (or $\mathcal{D}_d$) by using the procedure developed in [12] (see also [31]). This procedure is a form of *robustification* of the nominal system. Another possibility is to use a similarity transformation to look for a new representation of the system in which the transformed matrix $A_t \in \mathcal{D}$, as in the approach of [37].

Notice that we have concentrated on the determination of robust stability conditions without any additional requirements on the dynamics, such as optimality with respect to some index. These requirements can, however, be incorporated into the present analysis on the basis of, for example, [11], [32], and [34].

4. Illustrative examples. In this section we present some examples that appear in the literature and are described by the models treated in the previous sections. Robustness features of these systems are discussed from the perspective of our results.

Example 1: *Mathematical ecology.* Consider the generalized Lotka–Volterra model [35] given by the equations

$$\text{(26)} \qquad \dot{x}_i(t) = c_i x_i(t) + \sum_{j=1}^{n} a_{ij} x_i(t) x_j(t), \qquad i = 1, \ldots, n.$$

By assuming $x_i(t) \neq 0$, $i = 1, \ldots, n$, we can write

$$\{\dot{x}_i(t)/x_i(t)\} = c_i + \sum_{j=1}^{n} a_{ij} x_j(t), \qquad i = 1, \ldots, n,$$

$$\text{(27)} \qquad \Leftrightarrow \quad d(\ln x_i(t))/dt = c_i + \sum_{j=1}^{n} a_{ij} \exp(\ln x_j(t)), \qquad i = 1, \ldots, n.$$

By defining $y_i(t) := \ln x_i(t)$, we can rewrite (27) as

$$\dot{y}_i(t) = c_i + \sum_{j=1}^{n} a_{ij} f(y_j(t)), \qquad i = 1, \ldots, n,$$

and by assuming the existence of a positive equilibrium $y^* > 0$, i.e., a positive vector $y^* = (y_1^*, \ldots, y_n^*)^T$ satisfying $c_i + \sum_{j=1}^{n} a_{ij} f(y_j^*) = 0$, one can write the incremental model

$$\text{(28)} \qquad \dot{z}_i(t) = \sum_{j=1}^{n} a_{ij} g_j(z_j(t)), \qquad i = 1, \ldots, n,$$

where, for all $i, z_i(t) := y_i(t) - y_i^*$, $z^T(t) = (z_1(t), \ldots, z_n(t))$, and where $g_i : \xi \mapsto \exp(\xi + \ln x_i^*) - x_i^*$. Therefore, $g_i(0) = 0$ and $g_i(\xi) \cdot \xi > 0$ for $\xi \neq 0$. Consequently, (21) is a particular case of Persidskii's model (3) with the associated diagonal Liapunov function being given by

$$\text{(29)} \qquad V(z) = \sum_{i=1}^{n} p_i \int_0^{z_i} g_i(\tau) d\tau,$$

which corresponds to the "Volterra function" used in [13] and [14].

Computing $\dot{V}(z)$ along the trajectories of (28), one has

$$\dot{V}(z(t)) = g(z(t))^T(A^T D + DA)g(z(t)),$$

where $g(z)^T = [g_1(z_1), \ldots, g_n(z_n)]$ and $D = \frac{1}{2}\text{diag}(p_1, \ldots, p_n)$ and, provided $A \in \mathcal{D}_c$, $\dot{V}(x)$ is negative definite and the asymptotic stability of the equilibrium follows. □

From the discussion in §2, it is clear that diagonal stability of matrix $A = (a_{ij})$ implies the asymptotic stability of model (26). This condition, however, does not guarantee asymptotic stability with more general, nonlinear, and time-varying perturbations acting on this ecological system. To guarantee robustness under these circumstances, a more restrictive condition on the community matrix A (namely, $\overline{A} \in \mathcal{D}_c$; Theorem 2.5) is required, and this is the price that must be paid for robustness. This is also a way of interpreting Šiljak's result [35, Chap. 5], which gives the condition $\overline{A} \in \mathcal{D}_c$ to assure exponential stability of the equilibrium of (21).

Example 2: *Power systems, power-frequency control.* In power-system stability analysis and, more specifically, in power-frequency control, in which a multiple-machine system is considered, the Liapunov direct method has been used by several authors (see [30] and the references therein). In these analyses of power systems (considered without regulators), the "energy-type" Liapunov functions, which are widely used, are of the form

$$V(\omega, \delta) = \frac{1}{2}\sum_{i=1}^{n} M_i\omega_i^2 + \sum_{k=1}^{m}\int_0^{\delta_k} \psi_k(\tau)d\tau, \tag{30}$$

where $\delta^T = (\delta_1, \ldots, \delta_n)$ is the vector of incremental power angle differences between machines and $\omega^T = (\omega_1, \ldots, \omega_n)$ is the vector of incremental angular velocities; the function $\psi^T(\delta) = (\psi_1(\delta_1), \ldots, \psi_n(\delta_m))$ is composed of sinusoidal functions that (locally) satisfy the condition $\psi_i(\cdot) \in S_c$.

Clearly, (30) is a diagonal Liapunov function, and it is shown in [15] that in this case the incremental model for the multiple-machine power system can be expressed in the form

$$\begin{bmatrix} \dot{\omega} \\ \dot{\delta} \end{bmatrix} = A \begin{bmatrix} \omega \\ \psi(\delta) \end{bmatrix}, \tag{31}$$

where $A = (a_{ij})$ is a constant matrix. Consequently, (31) is a special case of (3) (for details and the definition of the $(n+m) \times (n+m)$ matrix A in (31), see [15]). Therefore, the asymptotic stability of the power system can be determined on the basis of the condition $A \in \mathcal{D}_c$, and *robust controllers* for this system can be designed on the basis of the same condition [15].

Example 3: *Digital filters with signal quantization.* The absence of limit cycles in zero-input digital filters with signal quantization is a problem that has been studied for several years (see [27] and the references therein). The special characteristics of signal quantization in state-space filters can be represented in terms of particular cases of the models considered in §2. The nonlinearities associated with quantization (truncation and rounding) and overflow operations are sector functions (class $S_d(k)$), and mathematical models associated with these types of digital filters are described by the equations

$$x_i(k+1) = \phi_i\left(\sum_{j=1}^{n} a_{ij}x_j(k)\right), \qquad i = 1, \ldots, n, \tag{32}$$

$$x_i(k+1) = \sum_{j=1}^{n} \phi_{ij}(a_{ij}x_j(k)), \qquad i = 1, \ldots, n, \tag{33}$$

where for all i, j, $\phi_i(\cdot) = \phi_{ij}(\cdot) = \phi(\cdot)$ is the scalar function associated with the particular type of quantization. Clearly, the above models are particular cases of (14) and (22), and (32) and (33) represent, respectively, the quantization operations performed after and before addition of the state variables in the digital filter.

Therefore, in studying the problem of robust stability (seeking conditions for the absence of self-sustained oscillations), Theorems 2.4 and 2.6 can be readily used, and the conditions $A \in \mathcal{D}_d$ and $|A| \in \mathcal{D}_d$ guarantee (respectively) this stability. These conditions have been derived in the literature: for model (32) the condition $A \in \mathcal{D}_d$ was derived in [28], and for model (33) the condition $|A| \in \mathcal{D}_d$ was derived in [19] (see also [22]) by using diagonal Liapunov functions.

To the illustrative examples presented above we can add many other examples in various fields in which diagonal stability is associated with robustness, e.g., the qualitatively stable (sign-stable) matrices that appear in mathematical economics and ecology [17], [26], which are also known to belong to class $\mathcal{D}_c$. Compartmental systems also belong to the class of systems that have a stable M-matrix structure. These systems have robust stability properties [24], and, again by the equivalence (24), these systems are also diagonally stable. In the field of chaotic asynchronous iterative computations, convergence of these (chaotic) computations is also ensured by the condition $|A| \in \mathcal{D}_d$, where A is an iteration matrix [7].

5. Concluding remarks. We have shown that diagonal Liapunov functions play an essential role in the robust stability problem of the linear autonomous continuous- and discrete-time systems $\dot{x}(t) = Ax(t)$ and $x(k+1) = Ax(k)$ subjected to some general classes of perturbations, and the many examples given support this conclusion.

Inspection of the proofs of the results above shows that for nondiagonal Liapunov functions it would be difficult or even impossible to perform the kind of majorizations necessary for the evaluation of bounds for $\dot{V}(\cdot)$ and $\Delta V(\cdot)$. This suggests that diagonal Liapunov functions are necessary if we restrict ourselves to the class of quadratic and integral-of-nonlinearity types of Liapunov functions, in the sense that one cannot do much better with nondiagonal Liapunov functions when the class of admissible perturbations is rich (i.e., when it includes time-varying sector-type nonlinearities and time-varying delays) and affects the nominal linear systems as modeled above ($\Sigma_c(t)$ and $\Sigma_d(k)$).

Acknowledgment. We would like to thank a semianonymous reviewer for comments and questions that helped to clarify the presentation of the results in the paper.

REFERENCES

[1] M. Araki and B. Kondo, *Stability and transient behavior of composite nonlinear systems*, IEEE Trans. Automat. Control, 17 (1972), pp. 537–541.

[2] G. P. Barker, A. Berman, and R. Plemmons, *Positive diagonal solutions to the Lyapunov equations*, Linear and Multilinear Algebra, 5 (1978), pp. 249–256.

[3] A. Berman and D. Hershkowitz, *Matrix diagonal stability and its implications*, SIAM J. Algebraic Discrete Meth., 4 (1983), pp. 377–382.

[4] S. P. Bhattacharyya, *Robust Stabilization Against Structured Perturbations*, Springer-Verlag, Berlin, 1987.

[5] S. Boyd and Q. Yang, *Structured and simultaneous Lyapunov functions for system stability problems*, Internat. J. Control, 49 (1989), pp. 2215–2240.

[6] D. Carlson, B. N. Datta, and C. R. Johnson, *A semi-definite Lyapunov theorem and the characterization of tridiagonal D-stable matrices*, SIAM J. Algebraic Discrete Meth., 3 (1982), pp. 293–304.

[7] D. Chazan and W. L. Miranker, *Chaotic relaxation*, Linear Algebra Appl., 2 (1969), pp. 190–222.

[8] P. Dorato, ed., *Robust Control*, IEEE Press, Washington, DC, 1987.

[9] M. Fiedler and V. Pták, *On matrices with non-positive off-diagonal elements and positive principal minors*, Czech. Math. J., 12 (1962), pp. 382–400.

[10] J. C. Geromel, *On the determination of a diagonal solution of the Liapunov equation*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 404–406.

[11] J. C. Geromel and J. J. Da Cruz, *On the robustness of optimal regulators for nonlinear discrete-time systems*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 703–710.

[12] J. C. Geromel and A. O. E. Santo, *On the robustness of linear continuous time dynamic systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 1136–1138.

[13] B. S. Goh, *Nonvulnerability of ecosystems in unpredictable environments*, Theoret. Population Biol., 10 (1976), pp. 83–95.

[14] ———, *Global stability in many-species systems*, Amer. Naturalist, 111 (1977), pp. 135–143.

[15] L. Hsu, L. A. Salgado, and E. Kaszkurewicz, *Structural properties in the stability problem of interconnected systems*, in Proc. 2nd International Federation of Automatic Control Symposium on Large Scale Systems, Toulouse, France, 1980, pp. 67–77.

[16] H. Hu, *An algorithm for scaling a matrix positive definite*, Linear Algebra Appl., 96 (1987), pp. 131–147.

[17] C. Jeffries, V. Klee, and P. V. D. Driessche, *When is a matrix sign stable?*, Canad. J. Math., 29 (1977), pp. 315–326.

[18] E. Kaszkurewicz, A. Bhaya, and D. D. Šiljak, *On the convergence of parallel asynchronous block-iterative computations*, Linear Algebra Appl., 131 (1990), pp. 139–160.

[19] E. Kaszkurewicz, A. Hermeto, and L. Hsu, *A result on stability of nonlinear discrete time systems and its application to recursive digital filters*, in Proc. 23rd IEEE Conference on Decision and Control, Las Vegas, NV, Institute of Electrical and Electronics Engineers, New York, 1984, pp. 100–102.

[20] E. Kaszkurewicz and L. Hsu, *A note on the absolute stability of nonlinear discrete-time systems*, Internat. J. Control, 40 (1984), pp. 867–869.

[21] ———, *On two classes of matrices with positive diagonal solutions to the Lyapunov equation*, Linear Algebra Appl., 59 (1984), pp. 19–27.

[22] ———, *Stability conditions for recursive digital filters with signal quantization*, Tech. Rep. EE-3, Dept. of Electrical Engineering, COPPE/UFRJ, Rio de Janeiro, Brazil, 1988.

[23] H. K. Khalil, *On the existence of positive diagonal P such that $PA + A'P < 0$*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 181–184.

[24] G. S. Ladde, *Cellular systems* II. *Stability of compartmental systems*, Math. Biosci., 30 (1976), pp. 1–21.

[25] M. Mansour, *Comment on stability of interval matrices*, Internat. J. Control, 46 (1987), p. 1845.

[26] J. S. Maybee and J. P. Quirk, *Qualitative problems in matrix theory*, SIAM Rev., 11 (1969), pp. 30–51.

[27] A. N. Michel and R. K. Miller, *Stability analysis of discrete-time interconnected systems via computer generated Lyapunov functions with applications to digital filters*, IEEE Trans. Circuits and Systems, CAS-32 (1985), pp. 737–753.

[28] W. L. Mills, C. T. Mullis, and R. Roberts, *Digital filter realizations without overflow oscillations*, IEEE Trans. Acoust. Speech Signal Process., ASSP-26 (1978), pp. 334–338.

[29] P. J. Moylan, *Matrices with positive principal minors*, Linear Algebra Appl., 17 (1977), pp. 53–58.

[30] M. A. Pai, *Power System Stability*, Elsevier North-Holland, Amsterdam, the Netherlands, 1981.

[31] R. V. Patel and M. Toda, *Quantitative measures of robustness for linear multivariable systems*, in Proc. Joint Automatic Control Conference, San Francisco, paper TP8, 1981.

[32] R. V. Patel, M. Toda, and S. Sridhar, *Robustness of linear quadratic state feedback designs in the presence of uncertainty*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 945–949.

[33] S. K. Persidskii, *Problem of absolute stability*, Automat. Remote Control, 12 (1969), pp. 1889–1895.

[34] M. E. Sezer and D. D. Šiljak, *Robust stability of discrete systems*, Internat. J. Control, 48 (1988), pp. 2055–2063.

[35] D. D. Šiljak, *Large Scale Dynamic Systems: Stability and Structure*, Elsevier North-Holland, Amsterdam, the Netherlands, 1978.

[36] ———, *Parameter space methods for robust control design: a guided tour*, IEEE Trans. Automat. Control, 34 (1989), pp. 674–688.

[37] R. K. Yedavalli, *Perturbation bounds for robust stability in linear state space models*, Internat. J. Control, 42 (1985), pp. 1507–1517.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 521–544, April 1993

SPECTRAL PROPERTIES OF PRECONDITIONED RATIONAL TOEPLITZ MATRICES: THE NONSYMMETRIC CASE*

TA-KANG KU† AND C.-C. JAY KUO†

Abstract. Various preconditioners for symmetric positive-definite (SPD) Toeplitz matrices in circulant matrix form have recently been proposed. The spectral properties of the preconditioned SPD Toeplitz matrices have also been studied. In this research, Strang's preconditioner S_N and our preconditioner K_N are applied to an $N \times N$ nonsymmetric (or nonhermitian) Toeplitz system $T_N \mathbf{x} = \mathbf{b}$. For a large class of Toeplitz matrices, it is proved that the singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ are clustered around unity except for a fixed number independent of N. If T_N is additionally generated by a rational function, the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ can be characterized directly. Let the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ be classified into the outliers and the clustered eigenvalues depending on whether they converge to 1 asymptotically. Then, the number of outliers depends on the order of the rational generating function, and the clustering radius is proportional to the magnitude of the last elements in the generating sequence used to construct the preconditioner. Numerical experiments are provided to illustrate our theoretical study.

Key words. Toeplitz matrix, preconditioned iterative method, rational generating function, nonsymmetric matrices

AMS(MOS) subject classifications. 65F10, 65F15

1. Introduction.

Research on preconditioning symmetric positive-definite (SPD) Toeplitz matrices with circulant matrices has been active recently [1], [6], [8], [9], [17]. In this research, we generalize Strang's preconditioner S_N [17] and our preconditioner K_N [9] to nonsymmetric (or nonhermitian) Toeplitz matrices. Let T_N be an $N \times N$ nonsymmetric Toeplitz matrix with elements $t_{i,j} = t_{i-j}$. The generalized Strang's preconditioner S_N is obtained by preserving N consecutive diagonals in T_N, i.e., diagonals with elements $t_n, 1 - M \leq n \leq N - M$, and using them to form a circulant matrix. One simple rule to determine M is to choose its value such that $|t_{N-M}| \approx |t_{1-M}|$. Note that half of the elements in T_N are not used in constructing S_N. The generalized preconditioner K_N is obtained from a $2N \times 2N$ circulant matrix in such a way that all elements in T_N are used, and is a circulant matrix itself (see §2). Since S_N and K_N are circulant, the matrix-vector products $S_N^{-1}\mathbf{v}$ and $K_N^{-1}\mathbf{v}$ can be conveniently computed via fast Fourier transform (FFT) with $O(N \log N)$ operations. The system of equations associated with the preconditioned Toeplitz matrix is then solved by iterative methods such as CGN (the conjugate gradient iteration applied to the normal equations) [7], GMRES (the generalized minimal residual) [15], and CGS (the conjugate gradient squared) [16].

The convergence rate of preconditioned iterative methods depends on the singular value or eigenvalue distribution of the preconditioned matrices [13]. The spectral properties of preconditioned SPD Toeplitz matrices have been widely studied. Chan and Strang [1], [5] proved that, for a symmetric Toeplitz with a positive generating function in the Wiener class, the preconditioned matrix has eigenvalues clustered around unity except a fixed number independent of N. If the Toeplitz is additionally

* Received by the editors April 16, 1991; accepted for publication (in revised form) September 4, 1991. This work was supported by the University of Southern California Faculty Research and Innovation Fund and by a National Science Foundation Research Initiation Award (ASC-9009323).

† Signal and Image Processing Institute and Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, California 90089-2564 (tkku@sipi.usc.edu and cckuo@sipi.usc.edu).

generated by a rational function, even stronger results were proved by Trefethen [19] and the authors [12]. In contrast, relatively few results for preconditioned nonsymmetric Toeplitz have been obtained so far [4], [11], [14], [21]. One such preconditioner was studied by R. Chan and Yeung [4], where they generalized T. Chan's preconditioner C_N to nonsymmetric Toeplitz matrices.

In this research, we examine the spectral properties of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ for nonsymmetric T_N in general, and nonsymmetric rational T_N in particular. The main results of our study are stated as follows. For a large class of general Toeplitz matrices, we prove that the singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$, or equivalently, the eigenvalues of $(S_N^{-1}T_N)^H(S_N^{-1}T_N)$ and $(K_N^{-1}T_N)^H(K_N^{-1}T_N)$, are clustered around unity except for a fixed number independent of N. If T_N is additionally generated by a rational function of order $(\alpha, \beta, \gamma, \delta)$, we are able to characterize the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ directly. We classify the eigenvalues of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ into two classes, i.e. the outliers and the clustered eigenvalues, depending on whether they converge to 1 asymptotically. Then, (i) the number of outliers is at most $\eta = 2\min(r, s)$ where $r = \max(\alpha, \beta)$ and $s = \max(\gamma, \delta)$; and (ii) the clustered eigenvalues are confined in a disk centered at 1 with radius ϵ, where the clustering radius ϵ is proportional to the magnitude of the last elements in the generating sequence used to construct the preconditioner.

With these spectral regularities, we can find appropriate preconditioned iterative methods to solve a nonsymmetric Toeplitz system efficiently. In particular, an $N \times N$ rational Toeplitz system $T_N\mathbf{x} = \mathbf{b}$ can be solved with $O(N \log N)$ operations since the number of iterations required for convergence is independent of the problem size N. To compare the performance of S_N and K_N, the $S_N^{-1}T_N$ and $K_N^{-1}T_N$ have the same number of outliers so that they converge in the same number of iterations asymptotically. However, the performances of S_N and K_N for finite N are determined by the clustering radii of the clustered eigenvalues as well. The magnitudes of the last elements used to construct S_N and K_N are $O(|t_{N-M}| + |t_{1-M}|)$ and $O(|t_N| + |t_{-N}|)$, respectively. Since $O(|t_N| + |t_{-N}|) \leq O(|t_{N-M}| + |t_{1-M}|)$ for large N, iterative methods with preconditioner K_N converges faster than with preconditioner S_N for solving rational Toeplitz systems. This is confirmed by numerical experiments. By the parallelism provided by FFT, the iterative methods with preconditioners in circulant matrix form is highly parallelizable, and the time complexity of the method can be reduced to $O(\log N)$ if $O(N)$ processors are used.

When T_N is a symmetric rational Toeplitz, we have $r = s$ and $t_N = t_{-N}$. Consequently, the number of outliers of $K_N^{-1}T_N$ is $\eta = 2r = 2\max(\alpha, \beta)$ and the clustering radius is $O(|t_N|)$. They reduce to the case given in [12]. Although the results derived in this paper can be viewed as a generalization of the results in [12], we want to point out that the approach adopted in this research is very different from that in [12] and the proof techniques are much more involved. For example, in characterizing the clustering radius of clustered eigenvalues of $K_N^{-1}T_N$ (or $S_N^{-1}T_N$) for symmetric T_N, the intertwining theorem of eigenvalues was exploited in [12]. However, such a theorem does not exist for nonsymmetric matrices so that we use perturbation theory for eigenvalues instead.

It is worthwhile to mention that there exists a preconditioner based on the minimum-phase LU factorization (MPLU) technique [11] which has a faster or comparable convergence rate than preconditioners S_N and K_N. However, Toeplitz preconditioners in circulant matrix form have two advantages over the MPLU preconditioner. First, the circulant preconditioning technique can be easily generalized to multidimensional Toeplitz systems. See [2], [3], and [10] for the two-dimensional case (block Toeplitz matrices). Second, the resulting preconditioned iterative method with preconditioners in circulant form is highly parallelizable while the MPLU preconditioner has to be implemented sequentially.

This paper is organized as follows. The construction of preconditioners S_N and K_N for nonsymmetric Toeplitz T_N is discussed in §2. We describe the singular value distribution of $K_N^{-1}T_N$ and $S_N^{-1}T_N$ for general Toeplitz in §3, and characterize the eigenvalue distribution of $K_N^{-1}T_N$ and $S_N^{-1}T_N$ for rational Toeplitz in §4 and §5, respectively. Numerical experiments are given in §6 to illustrate the theoretical study.

2. Constructions of Toeplitz preconditioners. Let T_m be a sequence of $m \times m$ nonsymmetric Toeplitz matrices with generating sequence t_n. Then,

$$T_N = \begin{bmatrix} t_0 & t_{-1} & \cdot & t_{-(N-2)} & t_{-(N-1)} \\ t_1 & t_0 & t_{-1} & \cdot & t_{-(N-2)} \\ \cdot & t_1 & t_0 & \cdot & \cdot \\ t_{N-2} & \cdot & \cdot & \cdot & t_{-1} \\ t_{N-1} & t_{N-2} & \cdot & t_1 & t_0 \end{bmatrix}.$$

Following the idea proposed by Strang [17], we construct the preconditioner S_N by preserving N consecutive diagonals in T_N and bringing them around to form a circulant matrix,

$$S_N = \begin{bmatrix} t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} & t_{N-M} & \cdot & t_2 & t_1 \\ t_1 & t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} & t_{N-M} & \cdot & t_2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & t_1 & t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} & t_{N-M} \\ t_{N-M} & \cdot & \cdot & t_1 & t_0 & t_{-1} & \cdot & t_{2-M} & t_{1-M} \\ t_{1-M} & t_{N-M} & \cdot & \cdot & t_1 & t_0 & t_{-1} & \cdot & t_{2-M} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ t_{-2} & \cdot & t_{1-M} & t_{N-M} & \cdot & \cdot & t_1 & t_0 & t_{-1} \\ t_{-1} & t_{-2} & \cdot & t_{1-M} & t_{N-M} & \cdot & \cdot & t_1 & t_0 \end{bmatrix}.$$

A simple rule of thumb to decide the value of M is to require $|t_{N-M}| \approx |t_{1-M}|$.

Generalizing the idea in [9], the preconditioner K_N is constructed based on a $2N \times 2N$ circulant matrix R_{2N},

$$R_{2N} = \begin{bmatrix} T_N & \triangle T_N \\ \triangle T_N & T_N \end{bmatrix},$$

where $\triangle T_N$ is determined by the elements of T_N to make R_{2N} circulant, i.e.,

$$\triangle T_N = \begin{bmatrix} 0 & t_{N-1} & \cdot & t_2 & t_1 \\ t_{-(N-1)} & 0 & t_{N-1} & \cdot & t_2 \\ \cdot & t_{-(N-1)} & 0 & \cdot & \cdot \\ t_{-2} & \cdot & \cdot & \cdot & t_{N-1} \\ t_{-1} & t_{-2} & \cdot & t_{-(N-1)} & 0 \end{bmatrix}.$$

This construction is motivated by the observation that the augmented circulant system,

$$\begin{bmatrix} T_N & \triangle T_N \\ \triangle T_N & T_N \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{b} \end{bmatrix},$$

is equivalent to $(T_N+\triangle T_N)\mathbf{x} = \mathbf{b}$ so that $(T_N+\triangle T_N)^{-1}\mathbf{b}$ can be computed efficiently via FFT and

$$K_N = T_N + \triangle T_N \tag{2.1}$$

can be used as a preconditioner for T_N. Note, however, that K_N itself is also circulant and can be inverted directly via N-point FFT rather than $2N$-point FFT.

3. Spectral properties of preconditioned Toeplitz. We assume that the generating sequence t_n satisfies the following two conditions:

$$\sum_{-\infty}^{\infty} |t_n| \le B_T < \infty, \tag{3.1}$$

$$|T(e^{i\theta})| = \left|\sum_{-\infty}^{\infty} t_n e^{-in\theta}\right| \ge \mu_T > 0 \quad \forall\theta. \tag{3.2}$$

According to the construction of preconditioners K_N and S_N, the function $T(e^{i\theta})$ describes the asymptotic eigenvalue distribution of preconditioners K_N and S_N but not necessarily the asymptotic eigenvalue distribution of Toeplitz matrix T_N (see Test Problem 5 in §6). Thus, the matrix-vector product $K_N^{-1}\mathbf{v}$ (or $S_N^{-1}\mathbf{v}$) can be completed successfully via FFT since magnitudes of eigenvalues of K_N^{-1} (or S_N^{-1}) are bounded due to (3.2). With conditions (3.1) and (3.2), the condition numbers of preconditioners K_N and S_N are bounded independent of N due to the following theorem.

THEOREM 1. *Let T_N be an $N \times N$ Toeplitz matrix with the corresponding generating sequence satisfying* (3.1) *and* (3.2). *The* $||(K_N K_N^H)^{-1}||_2$ *and* $||(S_N S_N^H)^{-1}||_2$ *are bounded for sufficiently large N.*

Proof. Since K_N is circulant, we have

$$K_N = F_N^H D_N F_N \quad \text{and} \quad K_N^H = F_N^H D_N^H F_N,$$

where F_N is the $N \times N$ unitary Fourier matrix with $N^{-1/2}e^{-i2\pi(m-1)(n-1)/N}$ as the (m,n) element and D_N a diagonal matrix formed by the eigenvalues of K_N. Thus, K_N, K_N^H, and $K_N K_N^H$ share the same eigenvectors, and the eigenvalues of $K_N K_N^H$ are

$$\lambda(K_N K_N^H) = \lambda(K_N)\lambda^*(K_N) = |\lambda(K_N)|^2.$$

Any eigenvalue of K_N belongs to the set of eigenvalues of R_{2N}, which are

$$\rho_n = \lambda_n(R_{2N}) = \sum_{k=-(N-1)}^{N-1} t_k e^{i2\pi kn/2N}, \qquad 1 \le n \le 2N.$$

It is clear that ρ_n is a partial sum of the infinite series $\sum_{-\infty}^{\infty} t_k e^{-ik\theta}$ with $\theta = -n\pi/N$. With (3.2), $|\rho_n| \ge \mu_T - \mu$, where μ can be made arbitrarily small by choosing sufficiently large N so that

$$||(K_N K_N^H)^{-1}||_2 \le \frac{1}{(\mu_T - \mu)^2} < \infty.$$

Similar arguments can be used to prove the boundness of $||(S_N S_N^H)^{-1}||_2$, and the proof is completed. □

The next theorem describes the clustering property of the singular values of $K_N^{-1}T_N$ and $S_N^{-1}T_N$.

THEOREM 2. *Let T_N be an $N \times N$ Toeplitz matrix with the generating sequence satisfying* (3.1) *and* (3.2). *For sufficiently large N, the singular values of the preconditioned matrices $K_N^{-1}T_N$ and $S_N^{-1}T_N$ are clustered around unity except for a fixed number independent of N.*

Proof. Note that the singular value of $K_N^{-1}T_N$ is equal to the square root of the corresponding eigenvalue of $(K_N^{-1}T_N)^H(K_N^{-1}T_N)$. Since $(K_N^{-1}T_N)^H(K_N^{-1}T_N)$ and $(K_NK_N^H)^{-1}(T_NT_N^H)$ are similar, the eigenvalues of $(K_NK_N^H)^{-1}(T_NT_N^H)$ are examined to understand the singular values of $K_N^{-1}T_N$. With the relation $K_N = T_N + \Delta T_N$, we have

$$\lambda[(K_NK_N^H)^{-1}(T_NT_N^H)] = 1 - \lambda[(K_NK_N^H)^{-1}(K_N\Delta T_N^H + \Delta T_NK_N^H - \Delta T_N\Delta T_N^H)].$$

Let us define

$$W_N = K_N\Delta T_N^H + \Delta T_NK_N^H - \Delta T_N\Delta T_N^H.$$

For an arbitrary positive integer q, we denote the corresponding central $(N-2q) \times (N-2q)$ diagonal block of $(K_NK_N^H)^{-1}$ and W_N by $\mathcal{K}_{N-2q}^{-1}$ and $\mathcal{W}_{N-2q}$, respectively. By the separation theorem (or intertwining theorem) of generalized eigenvalues [18], [20], there are at least $N-4q$ eigenvalues of $(K_NK_N^H)^{-1}W_N$ bounded by the minimum and the maximum eigenvalues of $\mathcal{K}_{N-2q}^{-1}\mathcal{W}_{N-2q}$.

Since $\mathcal{K}_{N-2q}^{-1}$ is a submatrix of the symmetric circulant matrix $(K_NK_N^H)^{-1}$,

$$||\mathcal{K}_{N-2q}^{-1}||_2 \le ||(K_NK_N^H)^{-1}||_2.$$

According to the definition of $\mathcal{W}_{N-2q}$,

$$\mathcal{W}_{N-2q} = \mathcal{K}\Delta\mathcal{T}^H + \Delta\mathcal{T}\mathcal{K}^H - \Delta\mathcal{T}\Delta\mathcal{T}^H,$$

where $\mathcal{K}$ and $\Delta\mathcal{T}$ are $(N-2q) \times N$ matrices formed by the central $(N-2q)$ rows of K_N and ΔT_N, respectively. It is easy to verify that, for $p = 1, \infty$,

$$||\mathcal{K}||_p \le 2 \sum_{n=-(N-1)}^{N-1} |t_n| \le 2 \sum_{n=-\infty}^{\infty} |t_n| \le 2B_T < \infty,$$

and

$$||\Delta\mathcal{T}||_p \le \sum_{n=q+1}^{N-1} (|t_n| + |t_{-n}|) \le \sum_{n=q+1}^{\infty} (|t_n| + |t_{-n}|) = \sigma(q).$$

Since $||A||_2 \le (||A||_1||A||_\infty)^{1/2}$ for an arbitrary matrix A, the above bounds also hold for $p = 2$. Similarly, we can argue that $||\mathcal{K}^H||_2 \le 2B_T < \infty$ and $||\Delta\mathcal{T}^H||_2 \le \sigma(q)$. Thus,

$$\begin{aligned}||\mathcal{W}_{N-2q}||_2 &\le ||\mathcal{K}||_2||\Delta\mathcal{T}^H||_2 + ||\Delta\mathcal{T}||_2||\mathcal{K}^H||_2 + ||\Delta\mathcal{T}||_2||\Delta\mathcal{T}^H||_2\\ &\le 4B_T\sigma(q) + \sigma^2(q).\end{aligned}$$

By using Theorem 1 and the fact that $\sigma(q)$ is smaller as q becomes larger due to (3.1), we conclude that for given ϵ there exist q and $\tilde{N}$ such that for all $N \ge \tilde{N}$,

$$||\mathcal{K}_{N-2q}^{-1}||_2||\mathcal{W}_{N-2q}||_2 \le ||(K_NK_N^H)^{-1}||_2||\mathcal{W}_{N-2q}||_2 \le \epsilon.$$

Hence, the eigenvalues of $(K_N K_N^H)^{-1}(T_N T_N^H)$ are confined in the interval $(1-\epsilon, 1+\epsilon)$ except at most $4q$ outlying eigenvalues. Similar arguments can be used to prove the spectral clustering property of the singular values of $S_N^{-1}T_N$. □

The solution of a Toeplitz system $T_N\mathbf{x} = \mathbf{b}$ can be determined by applying the CGN method to the preconditioned system $K_N^{-1}T_N\mathbf{x} = K_N^{-1}\mathbf{b}$ (or $S_N^{-1}T_N\mathbf{x} = S_N^{-1}\mathbf{b}$). With the clustering property on singular values of $K_N^{-1}T_N$ (or $S_N^{-1}T_N$), the CGN method converges superlinearly. A more detailed discussion about the convergence rate of the CGN method for solving Toeplitz systems were studied by Chan and Yeung [4]. When the generating function is additionally rational, we characterize the eigenvalues of the preconditioned matrices $K_N^{-1}T_N$ and $S_N^{-1}T_N$ directly. It will be detailed in the following sections.

4. Spectral properties of preconditioned rational Toeplitz $K_N^{-1}T_N$. The generating function of a sequence of Toeplitz matrices T_m is defined as

$$T(z) = \sum_{n=-\infty}^{\infty} t_n z^{-n}.$$

Let the generating function of T_N be of the form

$$T(z) = \frac{A(z^{-1})}{B(z^{-1})} + \frac{C(z)}{D(z)}, \tag{4.1}$$

where

$$\frac{A(z^{-1})}{B(z^{-1})} = \frac{a_0 + a_1 z^{-1} + \cdots + a_\alpha z^{-\alpha}}{1 + b_1 z^{-1} + \cdots + b_\beta z^{-\beta}}, \qquad \frac{C(z)}{D(z)} = \frac{c_0 + c_1 z + \cdots + c_\gamma z^\gamma}{1 + d_1 z + \cdots + d_\delta z^\delta}.$$

Note that $a_\alpha b_\beta c_\gamma d_\delta \neq 0$ and polynomials $A(z^{-1})$ and $B(z^{-1})$ (or $C(z)$ and $D(z)$) have no common factor. We call $T(z)$ a rational function of order $(\alpha, \beta, \gamma, \delta)$ and T_N a rational Toeplitz matrix. To simplify the notation, we define $r = \max(\alpha, \beta)$ and $s = \max(\gamma, \delta)$.

The spectral properties of $K_N^{-1}T_N$ can be determined from that of $T_N^{-1}\triangle T_N$ via

$$[\lambda(K_N^{-1}T_N)]^{-1} = \lambda(T_N^{-1}(T_N + \triangle T_N)) = 1 + \lambda(T_N^{-1}\triangle T_N). \tag{4.2}$$

The eigenvalues of $K_N^{-1}T_N$ clustered around 1 correspond to those of $T_N^{-1}\triangle T_N$ clustered around 0. We summarize the procedures in examining the spectral properties of $T_N^{-1}\triangle T_N$ as follows:

Step 1. Show that the $\triangle T_N$ is asymptotically equivalent to a low-rank Toeplitz matrix $\triangle F_N$ (Lemma 2).

Step 2. Study the rank of $\triangle F_N$ by transforming it to a matrix Q_F which has at most $d = r + s$ nonzero columns (Lemma 3).

Step 3. Show that the Q_F is asymptotically equivalent to a matrix $\overline{Q}_F$ which has at most $2\min(r, s)$ nonzero eigenvalues (Lemma 4).

Step 4. Use perturbation theory to determine the radius of the clustered eigenvalues of $T_N^{-1}\triangle T_N$ and $K_N^{-1}T_N$ (Lemmas 5 and 6 and Theorem 3).

The number of outliers of $K_N^{-1}T_N$, i.e., $2\min(r, s)$, is determined from Steps 1–3, and the clustering radius is determined from Step 4.

4.1. The number of outliers of $K_N^{-1}T_N$. Note that the sequence t_n can be recursively calculated for large $|n|$. This is stated as follows.

LEMMA 1. *The sequence t_n generated by* (4.1) *follows the recursions,*

$$(4.3)\quad \begin{aligned} t_{n+1} &= -(b_1 t_n + b_2 t_{n-1} + \cdots + b_\beta t_{n-\beta+1}), & n \geq r = \max(\alpha, \beta), \\ t_{n-1} &= -(d_1 t_n + d_2 t_{n+1} + \cdots + d_\delta t_{n+\delta-1}), & n \leq -s = -\max(\gamma, \delta). \end{aligned}$$

Proof. The proof is similar to the proof of Lemma 1 in [12]. □

Since elements t_n satisfy the recursion given in Lemma 1, we construct a low-rank Toeplitz matrix $\triangle F_N$ as

$$(4.4)\qquad \triangle F_N = F_{1,N} + F_{2,N},$$

where

$$F_{1,N} = \begin{bmatrix} t_N & t_{N-1} & \cdot & t_2 & t_1 \\ t_{N+1} & t_N & t_{N-1} & \cdot & t_2 \\ \cdot & t_{N+1} & t_N & \cdot & \cdot \\ t_{2N-2} & \cdot & \cdot & \cdot & t_{N-1} \\ t_{2N-1} & t_{2N-2} & \cdot & t_{N+1} & t_N \end{bmatrix}$$

and

$$F_{2,N} = \begin{bmatrix} t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} & t_{-(2N-1)} \\ t_{-(N-1)} & t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} \\ \cdot & t_{-(N-1)} & t_{-N} & \cdot & \cdot \\ t_{-2} & \cdot & \cdot & \cdot & t_{-(N+1)} \\ t_{-1} & t_{-2} & \cdot & t_{-(N-1)} & t_{-N} \end{bmatrix},$$

and where $t_n, n \geq r$ or $n \leq -s$, are recursively defined by (4.3). Due to the recursion given by (4.3), the ranks of $F_{1,N}$ and $F_{2,N}$ are bounded by r and s, respectively. Thus, the rank of $\triangle F$ is bounded by $d = r + s$. The following lemma shows that $\triangle T_N$ and $\triangle F_N$ are, in fact, asymptotically equivalent.

LEMMA 2. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in* (4.1) *with the corresponding generating sequence satisfying* (3.1) *and* (3.2). *The $\triangle T_N$ and $\triangle F_N$ are asymptotically equivalent.*

Proof. Let us denote the difference between $\triangle F_N$ and $\triangle T_N$ by

(4.5)

$$\triangle E_N = \triangle F_N - \triangle T_N = \begin{bmatrix} t_N + t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} & t_{-(2N-1)} \\ t_{N+1} & t_N + t_{-N} & t_{-(N+1)} & \cdot & t_{-(2N-2)} \\ \cdot & t_{N+1} & t_N + t_{-N} & \cdot & \cdot \\ t_{2N-2} & \cdot & \cdot & \cdot & t_{-(N+1)} \\ t_{2N-1} & t_{2N-2} & \cdot & t_{N+1} & t_N + t_{-N} \end{bmatrix}.$$

It can be easily verified that the l_1 and l_∞ norms of $\triangle E_N$ are both bounded by

$$(4.6)\qquad \tau_E = \sum_{n=N}^{2N-1} |t_n| + \sum_{n=-N}^{-(2N-1)} |t_n|.$$

Consequently, we have

$$||\triangle E_N||_2 \leq (||\triangle E_N||_1 ||\triangle E_N||_\infty)^{1/2} \leq \tau_E.$$

Since τ_E goes to zero as N goes to infinity due to (3.1), the proof is completed. □

Since $\triangle T_N$ is asymptotically equivalent to $\triangle F_N$ and the rank of $\triangle F_N$ is bounded by d, the number of outliers of $T_N^{-1}\triangle T_N$ (or $K_N^{-1}T_N$) is bounded by d, which is, however, not tight. We are able to determine a tighter bound by introducing another asymptotically equivalent matrix of $\triangle T_N$ (or $\triangle F_N$), which has only $2\min(r,s)$ nonzero eigenvalues in the following. This turns out to be the exact number of outliers actually observed in all our numerical experiments. To exploit the low-rank structure of $\triangle F_N$, we transform $\triangle F_N$ to

$$Q_F = \triangle F_N U_D L_B, \tag{4.7}$$

where U_D is an $N \times N$ upper triangular Toeplitz matrix with the first N coefficients in $D(z)$ as the first row, and L_B is an $N \times N$ lower triangular Toeplitz matrix with the first N coefficients in $B(z^{-1})$ as the first column. Note that since U_D and L_B are full-rank matrices, the Q_F and $\triangle F_N$ have the same rank. The structure of Q_F is described in the following lemma.

LEMMA 3. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in* (4.1) *with the corresponding generating sequence satisfying* (3.1) *and* (3.2). *The elements of Q_F are zeros except for the first s and the last r columns.*

Proof. Note that $F_{1,N}$ and $F_{2,N}$ are Toeplitz matrices with elements

$$(F_{1,N})_{i,j} = t_{N+i-j} \quad \text{and} \quad (F_{2,N})_{i,j} = t_{-N+i-j}.$$

The (i,j) elements of $F_{1,N}U_DL_B$ and $F_{2,N}U_DL_B$ are

$$\sum_{n=1}^{N}\sum_{m=1}^{N} t_{N+i-m}d_{n-m}b_{n-j} \quad \text{and} \quad \sum_{n=1}^{N}\sum_{m=1}^{N} t_{-N+i-m}d_{n-m}b_{n-j},$$

where $b_0 = 1$ ($d_0 = 1$) and $b_i = 0$ ($d_i = 0$) if the subscript i is not in the range $0 \le i \le \beta$ ($0 \le i \le \delta$). If $s < j \le N-r$, we can simplify the above summations as

$$\sum_{m'=0}^{\delta}\left(\sum_{n'=0}^{\beta} t_{N+i+m'-n'-j}b_{n'}\right)d_{m'} = 0$$

and

$$\sum_{n'=0}^{\beta}\left(\sum_{m'=0}^{\delta} t_{-N+i+m'-n'-j}d_{m'}\right)b_{n'} = 0,$$

where $m' = n - m$, $n' = n - j$, and the equalities are due to the recursion defined in (4.3). Thus, the elements of

$$Q_F = \triangle F_N U_D L_B = (F_{1,N} + F_{2,N})U_D L_B$$

are zeros except for the first s and the last r columns. □

Consequently, we decompose the complex N-tuple space C^N into two orthogonal complement subspaces,

$$\begin{aligned} \mathcal{R}(Q_F) &= \{\mathbf{v} \in C^N \mid \mathbf{v}_i = 0,\ s < i \le N-r\}, \\ \mathcal{N}(Q_F) &= \{\mathbf{v} \in C^N \mid \mathbf{v}_i = 0,\ 1 \le i \le s \ \text{ or } \ N-r < i \le N\}, \end{aligned} \tag{4.8}$$

with dimensions

$$\dim \mathcal{R}(Q_F) = d \quad \text{and} \quad \dim \mathcal{N}(Q_F) = N - d.$$

The subspace $\mathcal{N}(Q_F)$ is contained in the null space of Q_F. Let Q_{NW} denote the northwest $s \times s$ block in Q_F, and Q_{NE}, Q_{SW}, and Q_{SE} the corresponding corner blocks in Q_F with sizes $s \times r$, $r \times s$, and $r \times r$, respectively. By using the subspace decomposition (4.8), it is easy to see that the nonzero eigenvalues of Q_F only depend on the corresponding four corner blocks of Q_F, and are also the eigenvalues of the $d \times d$ matrix,

$$P_F = \begin{bmatrix} Q_{NW} & Q_{NE} \\ Q_{SW} & Q_{SE} \end{bmatrix}.$$

In other words, the rank of Q_F is the same as that of P_F.

The bounds for the elements of Q_{NW}, Q_{NE}, Q_{SW}, and Q_{SE} are summarized as follows:

$$\begin{aligned} &|(Q_{NW})_{i,j}| \le \tau_{NW}, && \tau_{NW} = O(|t_N| + |t_{-N}|), \\ &|(Q_{SE})_{i,j}| \le \tau_{SE}, && \tau_{SE} = O(|t_N| + |t_{-N}|), \\ &|(Q_{NE})_{i,j}| \le (F_{1,N}U_D L_B)_{i,N-r+j} + \tau_{NE}, && \tau_{NE} = O(|t_{-2N}|), \\ &|(Q_{SW})_{i,j}| \le (F_{2,N}U_D L_B)_{N-s+i,j} + \tau_{SW}, && \tau_{SW} = O(|t_{2N}|). \end{aligned} \tag{4.9}$$

To derive (4.9), recall that the (i,j) element of Q_F is

$$\sum_{n=1}^{N}\sum_{m=1}^{N} t_{N+i-m} d_{n-m} b_{n-j} + \sum_{n=1}^{N}\sum_{m=1}^{N} t_{-N+i-m} d_{n-m} b_{n-j},$$

which is bounded by

$$\sum_{m'=0}^{\delta}\sum_{n'=0}^{\beta} |t_{N+i+m'-n'-j}||d_{m'}||b_{n'}| + \sum_{m'=0}^{\delta}\sum_{n'=0}^{\beta} |t_{-N+i+m'-n'-j}||d_{m'}||b_{n'}|.$$

Since the elements of Q_{NW} are the same as those of Q_F with subscript (i,j), $i,j \le s$, they are bounded by

$$\tau_{NW} = \sum_{i=0}^{\beta} |b_i| \sum_{j=0}^{\delta} |d_j| \big(\max_{-(s+\beta)<n<s+\delta} |t_{N+n}| + \max_{-(s+\beta)<n<s+\delta} |t_{-N+n}| \big).$$

To determine the bound for $\sum_{i=0}^{\beta} |b_i|$, we factorize $B(z^{-1})$ as

$$B(z^{-1}) = (1 - r_1 z^{-1})(1 - r_2 z^{-1}) \cdots (1 - r_\beta z^{-1}).$$

A direct consequence of (3.1) is that all poles of $A(z^{-1})/B(z^{-1})$ should lie inside the unit circle, i.e., $|r_i| < 1$, $1 \le i \le \beta$, so that

$$|b_k| \le \binom{\beta}{k} (\max |r_i|)^k \le \binom{\beta}{k}, \quad \text{where} \quad \binom{\beta}{k} \equiv \frac{\beta!}{(\beta-k)!k!}.$$

Therefore, we obtain

$$\sum_{k=0}^{\beta} |b_k| \le \sum_{k=0}^{\beta} \binom{\beta}{k} \le 2^\beta.$$

Similarly, $\sum_{k=0}^{\delta} |d_k| \leq 2^{\delta}$ and thus, the elements of Q_{NW} are bounded by

$$\tau_{NW} = 2^{(\beta+\delta)}(|t_{N-s-\beta}| + |t_{-N+s+\delta}|) = O(|t_N| + |t_{-N}|),$$

where the last equality is due to the fact that, for large n, t_n can be approximated by

$$(4.10) \qquad t_n \approx cr_j^n, \quad \text{where} \quad |r_j| = \max_i |r_i|,$$

and where c is a constant. Similarly, we can prove that the elements of Q_{SE} are bounded by

$$\tau_{SE} = 2^{(\beta+\delta)}(|t_{N-r-\beta}| + |t_{-N+r+\delta}|) = O(|t_N| + |t_{-N}|).$$

The (i,j), $1 \leq i \leq s, 1 \leq j \leq r$, element of Q_{NE} is the sum of the $(i, N-r+j)$ elements of $F_{1,N}U_D L_B$ and $F_{2,N}U_D L_B$. It is straightforward to verify that the $(i, N-r+j)$ element of $F_{1,N}U_D L_B$ remains unchanged while that of $F_{2,N}U_D L_B$ is bounded by $\tau_{NE} = 2^{(\beta+\delta)}|t_{-2N+d+\delta}| = O(|t_{-2N}|)$ for sufficiently large N. Similarly, we can derive the bound for the elements in Q_{SW} as given by (4.9).

Thus, when N becomes asymptotically large, the P_F converges to

$$\overline{P}_F = \begin{bmatrix} 0 & \overline{Q}_{NE} \\ \overline{Q}_{SW} & 0 \end{bmatrix},$$

where $\overline{Q}_{NE}$ is the converged northeast $s \times r$ block in $F_{1,N}U_D L_B$ and $\overline{Q}_{SW}$ is the converged southwest $r \times s$ block in $F_{2,N}U_D L_B$. Since the ranks of $\overline{Q}_{NE}$ and $\overline{Q}_{SW}$ are both bounded by $\min(r,s)$, the rank of $\overline{P}_F$ is bounded by $\eta = 2\min(r,s)$.

Let us define a matrix $\overline{Q}_F$ by replacing the four corner blocks in Q_F with the corresponding blocks in $\overline{P}_F$. Then, we have

$$\begin{aligned} \tau_Q = ||Q_F - \overline{Q}_F||_p &= ||P_F - \overline{P}_F||_p \\ &\leq s\tau_{NW} + r\tau_{SE} + \max(r,s)(\tau_{NE} + \tau_{SW}) \\ &= O(|t_N| + |t_{-N}|), \end{aligned}$$

for $p = 1$ and ∞. The above bounds also hold for $p = 2$ because

$$||A||_2 \leq (||A||_1 ||A||_\infty)^{1/2}$$

for an arbitrary matrix A. Since τ_Q goes to zero as N goes to infinity due to (3.1), the asymptotic equivalence between Q_F and $\overline{Q}_F$ is established. This result is summarized in the following lemma.

LEMMA 4. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in* (4.1) *with the corresponding generating sequence satisfying* (3.1) *and* (3.2). *The Q_F and $\overline{Q}_F$ are asymptotically equivalent.*

Based on Lemmas 2–4, (4.2) and (4.7), $T_N^{-1}\triangle T_N$ is asymptotically equivalent to $T_N^{-1}\overline{Q}_F L_B^{-1} U_D^{-1}$ whose rank is bounded by $\eta = 2\min(r,s)$ and $K_N^{-1}T_N$ has at most η asymptotic eigenvalues not converging to one (outliers).

4.2. The clustering radius of $K_N^{-1}T_N$. We use perturbation theory to estimate the clustering radius of the $N - \eta$ clustered eigenvalues. Instead of examining the eigenvalues of $T_N^{-1}\triangle T_N$ directly, we study those of the similar matrix

$$G_N = L_B^{-1}U_D^{-1}T_N^{-1}\triangle T_N U_D L_B = L_B^{-1}U_D^{-1}T_N^{-1}Q_T,$$

where $Q_T = \Delta T_N U_D L_B$. Let us define

$$H_N = L_B^{-1} U_D^{-1} T_N^{-1} \overline{Q}_F.$$

It is clear that H_N has only d nonzero columns as $\overline{Q}_F$ (or Q_F). The G_N can be viewed as a matrix obtained from H_N by adding the perturbation matrix

$$\Delta G_N = G_N - H_N = L_B^{-1} U_D^{-1} T_N^{-1} (Q_T - \overline{Q}_F). \tag{4.11}$$

A bound of $||\Delta G_N||_2$ is given below so that we can estimate the clustering radius of the clustered eigenvalues by using perturbation theory for eigenvalues.

LEMMA 5. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in* (4.1) *with the corresponding generating sequence satisfying* (3.1) *and* (3.2). *Then, for sufficiently large N, the $||\Delta G_N||_2$ is bounded by $\epsilon = O(|t_N| + |t_{-N}|)$.*

Proof. We first study the 2-norm of $Q_T - \overline{Q}_F$, which is bounded by

$$||Q_T - \overline{Q}_F||_2 \leq ||Q_T - Q_F||_2 + ||Q_F - \overline{Q}_F||_2.$$

As shown in the proof of Lemma 4, the second term $||Q_F - \overline{Q}_F||_2$ is bounded by $\tau_Q = O(|t_N| + |t_{-N}|)$ while the first term $||Q_T - Q_F||_2$ is bounded by

$$||Q_T - Q_F||_2 \leq ||\Delta T_N - \Delta F_N||_2 ||U_D||_2 ||L_B||_2 = ||\Delta E_N||_2 ||U_D||_2 ||L_B||_2.$$

Recall from (4.6) that $||\Delta E_N||_2 \leq \sum_{n=N}^{2N-1} (|t_n| + |t_{-n}|)$. By using (4.10), we have

$$\sum_{n=N}^{2N-1} |t_n| \leq \sum_{n=N}^{\infty} |q_j r_j^n| = \frac{|t_N|}{1 - |r_j|} = M_B |t_N|, \quad \text{where} \quad M_B = \frac{1}{1 - |r_j|}.$$

Similarly, $\sum_{n=N}^{2N-1} |t_{-n}| \leq M_D |t_{-N}|$. Besides, $||L_B||_2 \leq \sum_{k=0}^{\beta} |b_k| \leq 2^\beta$ and $||U_D||_2 \leq \sum_{k=0}^{\delta} |d_k| \leq 2^\delta$. Thus, we obtain a bound for the first term, i.e.,

$$||Q_T - Q_F||_2 \leq 2^{(\beta+\delta)} (M_B |t_N| + M_D |t_{-N}|) = O(|t_N| + |t_{-N}|),$$

and conclude that

$$||Q_T - \overline{Q}_F||_2 \leq O(|t_N| + |t_{-N}|).$$

With (4.11), we have

$$||\Delta G_N||_2 \leq ||L_B^{-1}||_2 ||U_D^{-1}||_2 ||T_N^{-1}||_2 ||(Q_T - \overline{Q}_F)||_2.$$

Under the assumption that $||T_N^{-1}||_2$ is bounded by a constant c_T independent of N, we want to show that $||L_B^{-1}||_2$ and $||U_D^{-1}||_2$ are also bounded. Let us factorize $B(z^{-1})$ as

$$B(z^{-1}) = (1 - r_1 z^{-1})(1 - r_2 z^{-1}) \cdots (1 - r_\beta z^{-1}),$$

where we assume that all roots r_i are distinct for simplicity. By applying the isomorphism between the ring of the power series and the ring of semi-infinite lower (or upper) triangular Toeplitz matrices, the L_B and L_B^{-1} can be decomposed into the products

$$L_B = L_{r_1} L_{r_2} \cdots L_{r_\beta}, \qquad L_B^{-1} = L_{r_\beta}^{-1} \cdots L_{r_2}^{-1} L_{r_1}^{-1},$$

where $L_{r_i}, 1 \leq i \leq \beta$ is an $N \times N$ lower triangular Toeplitz matrix with $[1, -r_i, 0, \ldots, 0]^T$ as the first column. It can be easily verified that $L_{r_i}^{-1}$ is a lower triangular Toeplitz matrix with $[1, r_i, r_i^2, \ldots, r_i^{N-1}]^T$ as the first column. Therefore,

$$||L_{r_i}^{-1}||_p \leq \sum_{k=0}^{N-1} |r_i^k| \leq \sum_{k=0}^{\infty} |r_i^k| = \frac{1}{1-|r_i|}, \qquad p = 1, 2, \infty,$$

and

$$||L_B^{-1}||_2 \leq \prod_{i=1}^{\beta} ||L_{r_i}^{-1}||_2 \leq \prod_{i=1}^{\beta} \frac{1}{1-|r_i|} = c_B.$$

Similar arguments can be used to prove that $||U_D^{-1}||_2 \leq c_D$. Finally, we have

$$||\triangle G_N||_2 \leq \epsilon \equiv c_B c_D c_T ||(Q_T - \overline{Q}_F)||_2 = O(|t_N| + |t_{-N}|). \tag{4.12}$$

The proof is completed. □

Let us denote the rank of $H_N = L_B^{-1} U_D^{-1} T_N^{-1} \overline{Q}_F$ by $\tilde{\eta}$. Clearly, $\tilde{\eta} \leq \eta = 2\min(r, s)$. We arrange the eigenvalues of H_N in a descending order so that $|\lambda_n| \geq |\lambda_{n+1}|$ ($\lambda_n = 0$ for $\tilde{\eta} < n \leq N$), and denote the corresponding normalized right-hand and left-hand eigenvectors by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N$, respectively. Besides, vectors $\mathbf{x}_n$ with $\tilde{\eta} < n \leq N$ are chosen to be othorgonal. The complex N-tuple space is decomposed into the row and the null spaces of H_N,

$$\mathrm{Row}(H_N) = \mathrm{span}\{\mathbf{x}_n, n \leq \tilde{\eta}\}, \qquad \mathrm{Null}(H_N) = \mathrm{span}\{\mathbf{x}_n, \tilde{\eta} < n \leq N\}.$$

Since $G_N = H_N + \triangle G_N$ and $||\triangle G_N||_2 \leq \epsilon$, the eigenvalues and the right-hand eigenvectors of G_N are denoted by $\lambda_n(\epsilon)$ and $\mathbf{x}_n(\epsilon)$, respectively. By using results from perturbation theory for repeated eigenvalues [20], the eigenvectors $\mathbf{x}_n(\epsilon)$ with $\tilde{\eta} < n \leq N$ must take the form

$$\mathbf{x}_n(\epsilon) = \sum_{m=1}^{\tilde{\eta}} \frac{\xi_{mn}}{(\lambda_n - \lambda_m)s_m} \mathbf{x}_m + \sum_{m=\tilde{\eta}+1}^{N} g_{mn}\mathbf{x}_m + O(\epsilon^2), \tag{4.13}$$

where $\xi_{mn} = \mathbf{y}_m^H \triangle G_N \mathbf{x}_n$, $\lambda_n = 0$, $s_m = \mathbf{y}_m^H \mathbf{x}_m$ and $g_{nn} = 1$. Due to the construction, we know that

$$||\mathbf{x}_n(\epsilon)||_2 \geq ||\mathbf{x}_n||_2 = 1. \tag{4.14}$$

The factor $|\xi_{mn}|$ is bounded by

$$|\xi_{mn}| = |\mathbf{y}_m^H \triangle G_N \mathbf{x}_n| \leq ||\mathbf{y}_m||_2 ||\triangle G_N||_2 ||\mathbf{x}_n||_2 \leq \epsilon.$$

The $|s_m^{-1}|, 1 \leq m \leq \tilde{\eta}$, is also bounded as given in the following lemma.

LEMMA 6. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in* (4.1) *with the corresponding generating sequence satisfying* (3.1) *and* (3.2). *Then, the $|s_m^{-1}|, 1 \leq m \leq \tilde{\eta}$, of H_N is bounded by a constant independent of N.*

Proof. The eigenvalues λ and the right-hand eigenvectors $\mathbf{x}$ of H_N satisfy

$$L_B \overline{Q}_F \mathbf{x} = \lambda L_B T_N U_D L_B \mathbf{x}.$$

Since the elements of $\overline{Q}_F$ are zeros except the first s and the last r columns, so are the elements of $L_B\overline{Q}_F$. Thus, the nonzero eigenvalues of H_N only depend on the northwest $s \times s$, northeast $s \times r$, southwest $r \times s$, and southeast $r \times r$ blocks of $L_B\overline{Q}_F$ and $L_BT_NU_DL_B$. The boundness of $|s_m^{-1}|, 1 \leq m \leq \tilde{\eta}$, is guaranteed if the elements of the four corner blocks of $L_B\overline{Q}_F$ and $L_BT_NU_DL_B$ remain unchanged for sufficiently large N.

By using the band structure of L_B and the special structure of $\overline{Q}_F$, it is straightforward to verify that the four blocks of $L_B\overline{Q}_F$ remain unchanged for large N. Next, we examine the matrix $L_BT_NU_DL_B$. By using (4.1) and the isomorphism between the ring of the power series and the ring of the semi-infinite lower (or upper) triangular Toeplitz matrices, we can express T_N as

$$T_N = L_AL_B^{-1} + U_CU_D^{-1},$$

where L_A is an $N \times N$ lower triangular Toeplitz matrix with the first N coefficients in $A(z^{-1})$ as the first column, and U_C is an $N \times N$ upper triangular Toeplitz matrix with the first N coefficients in $C(z)$ as the first row. Then, we have

$$L_BT_NU_DL_B = L_AU_DL_B + L_BU_CL_B,$$

whose four corner blocks remain unchanged for large N. Thus, λ_m and $s_m = \mathbf{y}_m^H\mathbf{x}_m$ with $1 \leq m \leq \tilde{\eta}$ do not change with N, when N becomes sufficiently large. □

Let $\mathbf{v}_n(\epsilon)$ be the normalized vector of $\mathbf{x}_n(\epsilon)$,

$$\mathbf{v}_n(\epsilon) = \frac{\mathbf{x}_n(\epsilon)}{||\mathbf{x}_n(\epsilon)||_2},$$

which can be decomposed as

$$\mathbf{v}_n(\epsilon) = \mathbf{v}_{\mathrm{N}}(\epsilon) + \mathbf{v}_{\mathrm{R}}(\epsilon),$$

where $\mathbf{v}_{\mathrm{N}}(\epsilon) \in \mathrm{Null}(H_N)$ and $\mathbf{v}_{\mathrm{R}}(\epsilon) \in \mathrm{Row}(H_N)$. The magnitude of $\lambda_n(\epsilon)$, $\tilde{\eta} < n \leq N$, of G_N is approximated by

$$|\lambda_n(\epsilon)| = ||G_N\mathbf{v}_n(\epsilon)||_2 = ||H_N\mathbf{v}_{\mathrm{R}}(\epsilon) + \triangle G_N\mathbf{v}_n(\epsilon)||_2.$$

By using (4.12)–(4.14), we obtain that

$$\begin{aligned}\max_{\tilde{\eta}<n\leq N} |\lambda_n(\epsilon)| &\leq \max_{\tilde{\eta}<n\leq N} ||H_N\mathbf{v}_{\mathrm{R}}(\epsilon)||_2 + \max_{\tilde{\eta}<n\leq N} ||\triangle G_N\mathbf{v}_n(\epsilon)||_2 \\ &\leq \sum_{m=1}^{\tilde{\eta}} \frac{||\xi_{mn}H_N\mathbf{x}_m||_2}{|\lambda_m s_m|\ ||\mathbf{x}_n(\epsilon)||_2} + ||\triangle G_N||_2 \\ &\leq \sum_{m=1}^{\tilde{\eta}} \frac{\epsilon}{|s_m|} + \epsilon = \epsilon_K \\ &= O(|t_N| + |t_{-N}|)\end{aligned}$$

for sufficiently large N. The above analysis is concluded in the following theorem.

THEOREM 3. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in* (4.1) *with the corresponding generating sequence satisfying* (3.1) *and* (3.2). *For sufficiently large N, the preconditioned Toeplitz matrix $K_N^{-1}T_N$ has the following two properties:*

P1. *The number of outliers is at most $\eta = 2\min(r,s)$.*

P2. *There are at least $N-\eta$ eigenvalues confined in the disk centered at* 1 *with radius ϵ_K, where*

$$\epsilon_K = O(|t_N| + |t_{-N}|).$$

5. Spectral properties of preconditioned rational Toeplitz $S_N^{-1}T_N$. The preconditioned Toeplitz matrix $S_N^{-1}T_N$ has similar spectral properties as $K_N^{-1}T_N$. The number of outliers of $S_N^{-1}T_N$ can be obtained by proving that $\triangle S_N = S_N - T_N$ and $\triangle F_N$ given by (4.4) are asymptotically equivalent.

LEMMA 7. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in* (4.1) *with the corresponding generating sequence satisfying* (3.1) *and* (3.2). *$S_N^{-1}T_N$ has asymptotically at most $\eta = 2\min(r,s)$ eigenvalues not converging to* 1.

Proof. Let us define $\triangle S_N = S_N - T_N$, and express the difference between $\triangle F_N$ in (4.4) and $\triangle S_N$ as

$$\triangle F_N - \triangle S_N = E_{1,N} + E_{2,N},$$

where $E_{1,N}$ and $E_{2,N}$ are $N \times N$ Toeplitz matrices with elements

$$(E_{1,N})_{i,j} = \begin{cases} t_{N+i-j}, & -(M-1) \le i-j \le N-1, \\ t_{i-j}, & -(N-1) \le i-j \le -M, \end{cases}$$

and

$$(E_{2,N})_{i,j} = \begin{cases} t_{i-j}, & N-(M-1) \le i-j \le N-1, \\ t_{i-j-N}, & -(N-1) \le i-j \le N-M, \end{cases}$$

respectively. By using similar arguments in deriving Lemma 2, we can prove that $\triangle S_N$ and $\triangle F_N$ are asymptotically equivalent. Since $\triangle F_N$ is asymptotically equivalent to the matrix $\overline{Q}_F L_B^{-1} U_D^{-1}$ with rank $\tilde{\eta} \le \eta = 2\min(r,s)$ as described in Lemma 4, the proof is completed. □

Similar arguments used in §4.2 can be applied to derive the following theorem.

THEOREM 4. *Let T_N be an $N \times N$ Toeplitz matrix generated by $T(z)$ in* (4.1) *with the corresponding generating sequence satisfying* (3.1) *and* (3.2). *For sufficiently large N, the preconditioned Toeplitz matrix $S_N^{-1}T_N$ has the following two properties:*

P1. *The number of outliers is at most $\eta = 2\min(r,s)$.*

P2. *There are at least $N-\eta$ eigenvalues confined in the disk centered at* 1 *with radius ϵ_S, where*

$$\epsilon_S = O(|t_{N-M}| + |t_{1-M}|).$$

6. Numerical results. Five test problems, including both rational and nonrational T_N, are used to illustrate the above analysis. For the nonsymmetric Toeplitz system $T_N\mathbf{x} = \mathbf{b}$ to be solved, we choose $\mathbf{b} = (1,\ldots,1)^T$ and zero initial guesses in the first four experiments. Without further specification, M is chosen such that $|t_{N-M}| \approx |t_{1-M}|$ to construct preconditioner S_N. We use the first test problem, which is generated by a nonrational function, to examine the clustering effect of singular values. Test Problems 2–4 are generated by rational functions so that the number of outliers and the clustering radius can be observed, which confirm the theoretical

TABLE 1
Number of iterations required for the CGN *method.*

N	T_N	S_N	K_N
32	24	12	9
64	33	15	11
128	49	17	13

TABLE 2
Number of iterations required for the CGS *method.*

N	T_N	S_N	K_N
32	15	7	9
64	21	8	10
128	26	9	10

results developed in §4 and §5. One underdetermined Toeplitz system (singular T_N) is also given in Test Problem 5 with appropriate vector **b**.

TEST PROBLEM 1. Nonrational T_N. Let T_N be a Toeplitz matrix with generating sequence

$$t_n = \begin{cases} 1/\log(2-n), & n \le -1, \\ 1/\log(2-n) + 1/(1+n), & n = 0, \\ 1/(1+n), & n \ge 1. \end{cases}$$

The singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ are plotted in Fig. 1(a) for $N = 32$, 64, and 128. Both $S_N^{-1}T_N$ and $K_N^{-1}T_N$ have clustered singular values. The eigenvalues of $K_N^{-1}T_N$ with $N = 32$ are plotted in Fig. 1(b). It is clear that the eigenvalues possess a certain clustering property. We apply both the CGN and CGS methods to solve the preconditioned Toeplitz system $P_N^{-1}T_N\mathbf{x} = P_N^{-1}\mathbf{b}$. The numbers of iterations required for the CGN and CGS methods to achieve $||\mathbf{b} - T_N\mathbf{x}||_2 < 10^{-12}$ are summarized in Tables 1 and 2, respectively. The case without preconditioning, denoted by T_N, is also included for comparison. The use of preconditioners does accelerate the convergence rate of these iterative methods. The number of iterations required for S_N and K_N increase slightly as N becomes large. The K_N performs better than S_N in the CGN method. However, their performances are comparable for the CGS method. Note also that it requires more iterations for the CGN method to converge than the CGS method. Since the operation counts inside each iteration of the CGN and CGS methods are approximately equal, the CGS method performs better for this test problem. However, the CGN method may have a more robust convergence behavior for general problems. We refer to [13] for a more detailed comparison of these two nonsymmetric iterative algorithms.

Although the necessary conditions for convergence of the CGS method are not clear yet, there exists a direct relation between the convergence rate and the eigenvalue distribution of the iteration matrix [11], [13]. We will only present results of the CGS method for remaining test problems to verify the theoretic results about the eigenvalue distributions of the preconditioned rational Toeplitz matrices.

TEST PROBLEM 2. Rational T_N with $(r, s) = (1, 1)$. The generating function of T_N is chosen to be

$$T(z) = \frac{1 + 0.7z^{-1}}{1 - 0.9z^{-1}} + \frac{1 - 0.8z}{1 + 0.7z}.$$

To show that the simple rule for choosing M, i.e., $|t_{N-M}| \approx |t_{1-M}|$, does provide a better spectral clustering property and a better convergence rate for $S_N^{-1}T_N$, two

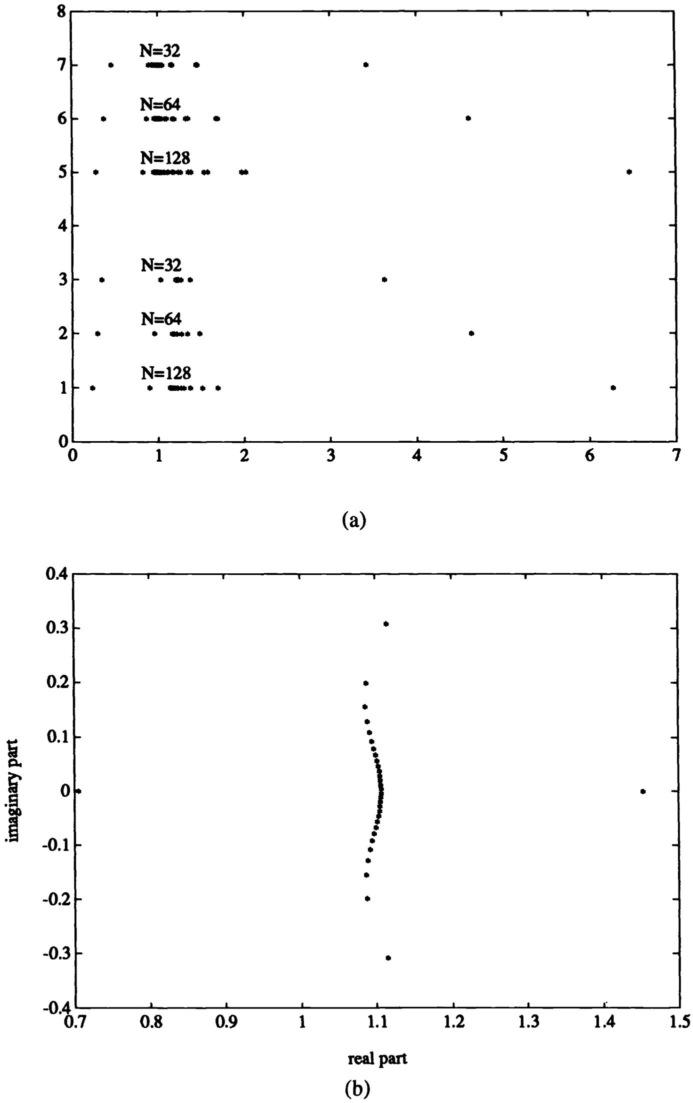


FIG. 1. (a) *The singular value distribution of* $S_N^{-1}T_N$ *and* $K_N^{-1}T_N$, *and* (b) *the eigenvalue distribution of* $K_N^{-1}T_N$ *for Test Problem* 1.

preconditioners S_N and $\tilde{S}_N$ are constructed. The S_N is constructed with M such that $|t_{N-M}| \approx |t_{1-M}|$ while the $\tilde{S}_N$ is constructed with $M = \lceil N/2 \rceil$. The eigenvalues of T_N, $\tilde{S}_N^{-1}T_N$, $S_N^{-1}T_N$ and $K_N^{-1}T_N$ with $N = 32$ are plotted in Figs. 2(a)–(d). All preconditioned Toeplitz matrices have eigenvalues clustered around 1 except $2 = 2\min(r,s)$ outliers. The $K_N^{-1}T_N$ has the best clustering effect, and the eigenvalues of $S_N^{-1}T_N$ are more closely clustered than those of $\tilde{S}_N^{-1}T_N$. The sums of magnitudes of the last elements in constructing S_N and K_N and the corresponding clustering radii are listed in Table 3. They are of approximately the same order, as stated in Theorems 3 and 4.

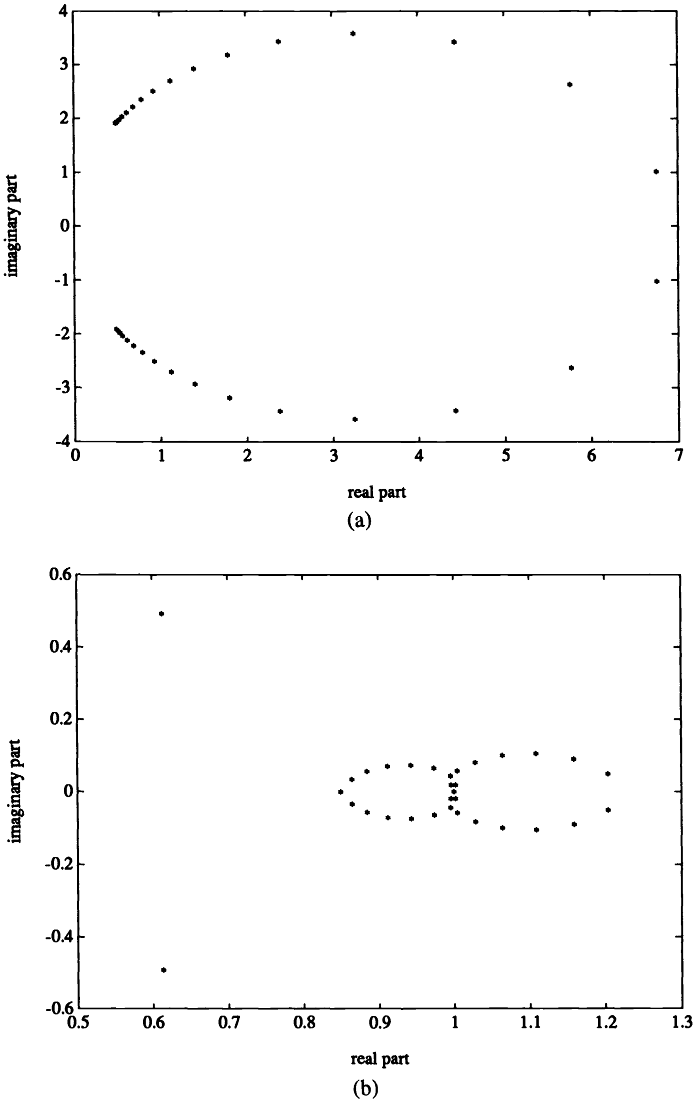
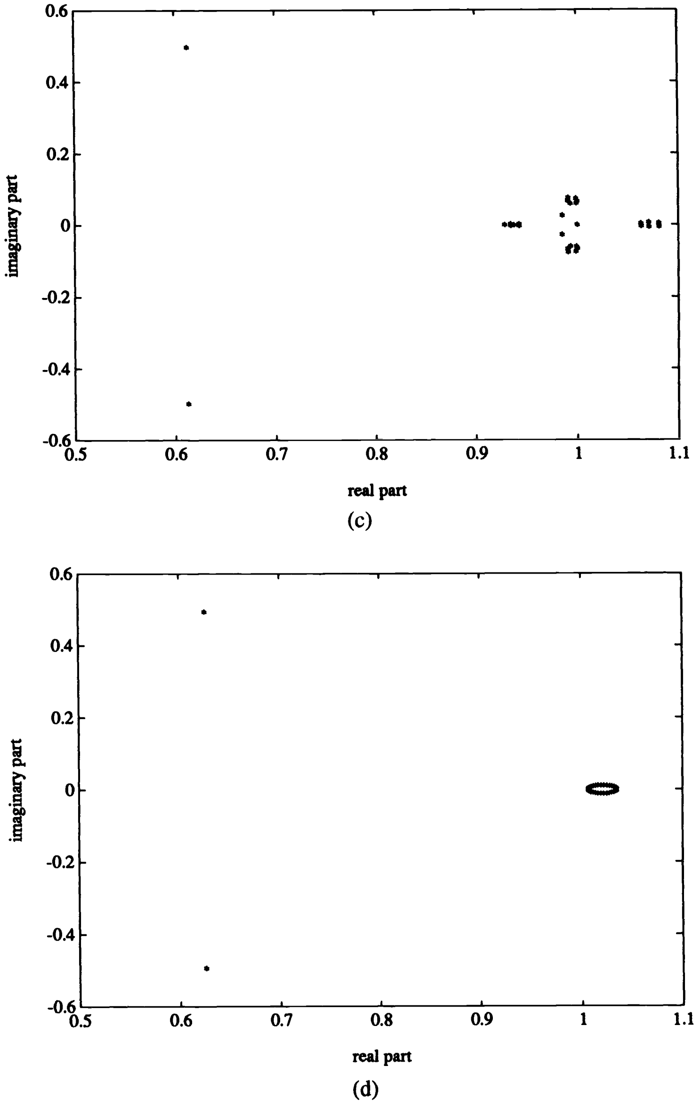


FIG. 2. *The eigenvalue distribution of* (a) T_N *and* (b) $\tilde{S}_N^{-1}T_N$ *for Test Problem* 2.

The convergence history of the CGS method with various preconditioners is plotted in Fig. 3 with $N = 32$. The convergence rate of the CGS method without preconditioning (the curve denoted by T_N) is very slow. This phenomenon is not surprising by examining the eigenvalue distribution given in Fig. 2(a). Preconditioning improves the convergence behavior dramatically. It is clear that K_N performs the best, while S_N performs better than $\tilde{S}_N$.

TEST PROBLEM 3. Rational T_N with $(r, s) = (3, 1)$. The generating function of



(c)

(d)

FIG. 2. *The eigenvalue distribution of* (c) $S_N^{-1}T_N$ *and* (d) $K_N^{-1}T_N$ *for Test Problem* 2.

T_N is chosen to be

$$T(z) = \frac{(1+0.5z^{-1})(1+0.7z^{-1})}{(1-0.4z^{-1})(1-0.6z^{-1})(1-0.8z^{-1})} + \frac{1+0.8z}{1+0.9z}.$$

The eigenvalues of T_N, $S_N^{-1}T_N$, and $K_N^{-1}T_N$ with $N = 64$ are plotted in Figs. 4(a)–(c). It is clear that $K_N^{-1}T_N$ has $2 = 2\min(r, s)$ outliers. The outliers of $S_N^{-1}T_N$ are not easy to identify for this case. However, two outliers can be observed more easily for larger N. Besides, the eigenvalues of $K_N^{-1}T_N$ are more closely clustered than those of $S_N^{-1}T_N$. We list in Table 4 the sums of magnitudes of the last elements

TABLE 3
Clustering radii ϵ of preconditioners S_N and K_N for Test Problem 2.

N	ϵ_S	$\lvert t_{N-M}\rvert + \lvert t_{1-M}\rvert$	ϵ_K	$\lvert t_{N-1}\rvert + \lvert t_{1-N}\rvert$
32	8.2×10^{-2}	2.8×10^{-1}	3.5×10^{-2}	6.8×10^{-2}
64	4.6×10^{-2}	2.1×10^{-2}	1.2×10^{-3}	2.3×10^{-3}
128	3.3×10^{-5}	1.1×10^{-4}	1.4×10^{-6}	2.7×10^{-6}

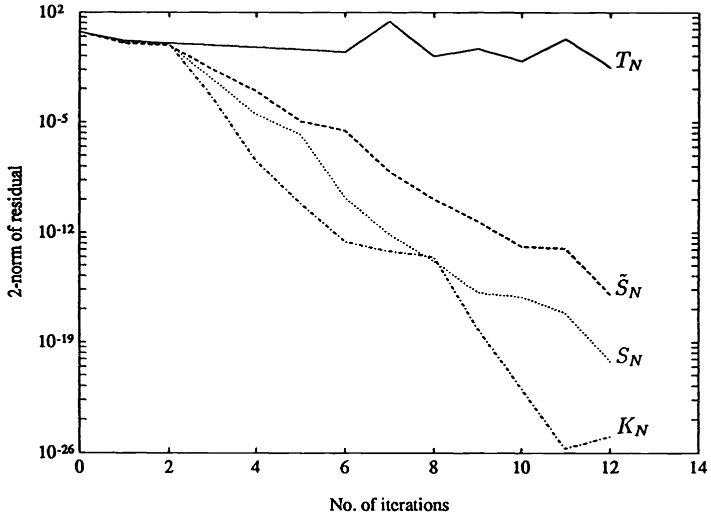


FIG. 3. *Convergence history of the* CGS *method for Test Problem* 2.

in constructing S_N and K_N and the corresponding clustering radii. The convergence history of the CGS method with $N = 64$ is plotted in Fig. 5. We observe that the CGS method without preconditioning does not converge and that the CGS method with preconditioners K_N and S_N converges in 4 and 6 iterations, respectively. This seems to suggest that the use of preconditioners does not only accelerate the convergence rate by providing better spectral properties, but also improves the convergence of nonsymmetric iterative algorithms by making the preconditioned matrix more close to normal.

TEST PROBLEM 4. Rational triangular T_N with $(r, s) = (1, 0)$. The generating function of T_N is chosen to be

$$T(z) = \frac{1 - 0.7z^{-1}}{1 + 0.5z^{-1}}.$$

Since there are only N nonzero elements in T_N, we can make S_N the same as K_N. The eigenvalues of $K_N^{-1}T_N$ with $N = 32$ are plotted in Fig. 6(a). We see that all eigenvalues are clustered around 1 with radius $\epsilon_K = O(|t_N|) = 10^{-9}$. This is consistent with Theorem 3, which predicts that $K_N^{-1}T_N$ has $0 = 2\min(r, s)$ outliers. The convergence history of the CGS method with $N = 32$ is plotted in Fig. 6(b). The CGS method with preconditioner K_N converges in two iterations while the CGS method without preconditioning does not converge.

TEST PROBLEM 5. Ill-conditioned T_N. Let T_N be the $N \times N$ Toeplitz matrix

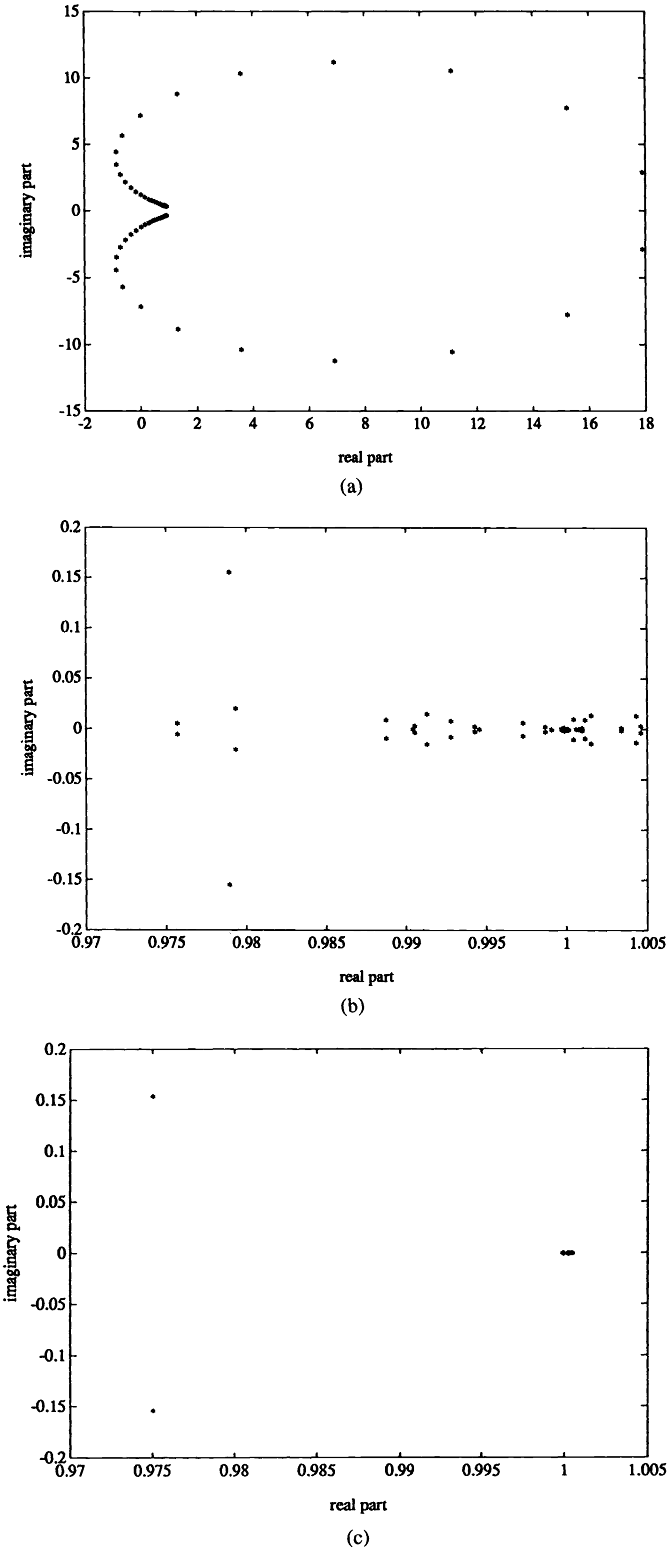


FIG. 4. *The eigenvalue distribution of* (a) T_N, (b) $S_N^{-1}T^N$, *and* (c) $K_N^{-1}T_N$ *for Test Problem* 3.

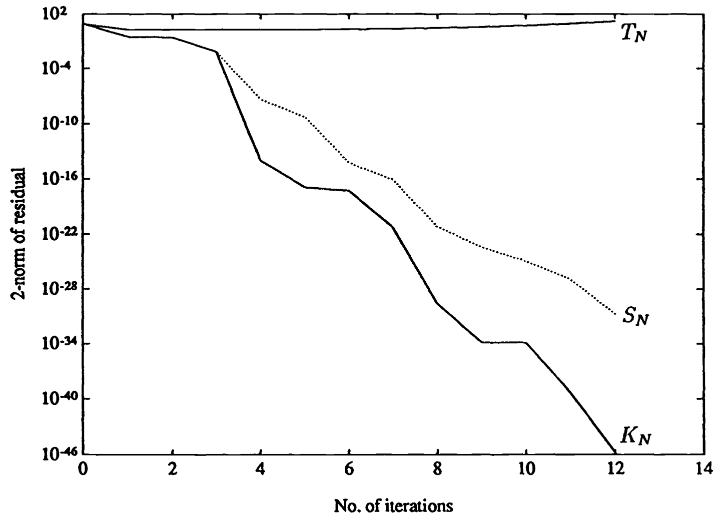


FIG. 5. *Convergence history of the* CGS *method for Test Problem* 3.

TABLE 4
Clustering radii ϵ of preconditioners S_N and K_N for Test Problem 3.

N	ϵ_S	$\lvert t_{N-M}\rvert + \lvert t_{1-M}\rvert$	ϵ_K	$\lvert t_{N-1}\rvert + \lvert t_{1-N}\rvert$
32	1.7×10^{-1}	1.4×10^{-1}	6.1×10^{-2}	2.8×10^{-2}
64	2.7×10^{-2}	1.3×10^{-2}	5.1×10^{-4}	1.6×10^{-4}
128	1.7×10^{-3}	1.6×10^{-4}	5.8×10^{-7}	1.7×10^{-7}

generated by polynomial $T(z) = \varepsilon + z^{-1}$ with $\varepsilon < 1$, i.e.,

$$T_N = \begin{bmatrix} \varepsilon & 0 & 0 & 0 & \cdot & 0 & 0 \\ 1 & \varepsilon & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & \varepsilon & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \varepsilon & 0 & 0 \\ 0 & \cdot & \cdot & 0 & 1 & \varepsilon & 0 \\ 0 & 0 & \cdot & 0 & 0 & 1 & \varepsilon \end{bmatrix}.$$

Matrix T_N is nearly singular for large N and $||T_N^{-1}||$ increase with N at the rate $O(\varepsilon^{-N})$.

The preconditioner K_N $(= S_N)$ is

$$K_N = \begin{bmatrix} \varepsilon & 0 & 0 & 0 & \cdot & 0 & 1 \\ 1 & \varepsilon & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & \varepsilon & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \varepsilon & 0 & 0 \\ 0 & \cdot & \cdot & 0 & 1 & \varepsilon & 0 \\ 0 & 0 & \cdot & 0 & 0 & 1 & \varepsilon \end{bmatrix},$$

and $||K_N^{-1}||$ is bounded for large N. Since the rank of the matrix $K_N - T_N$ is one, $K_N^{-1}T_N$ has $N - 1$ eigenvalues repeated at 1 and our theoretic results still hold.

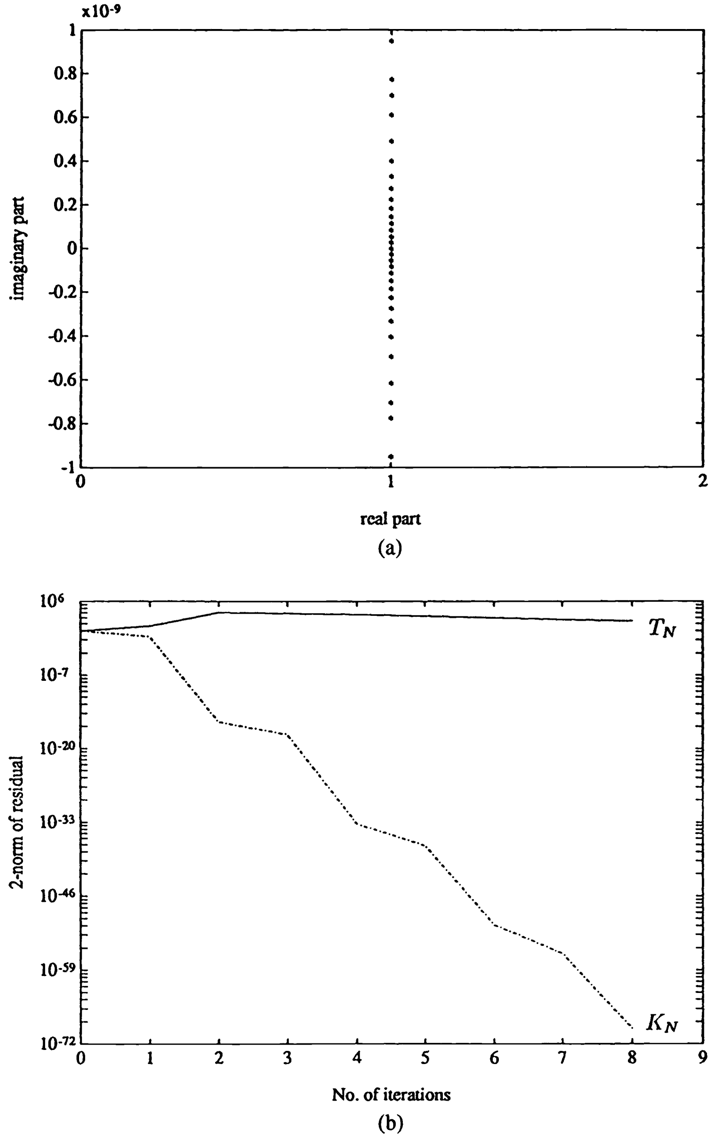


FIG. 6. (a) *The eigenvalue distribution of* $K_N^{-1}T_N$ *and* (b) *the convergence history of the* CGS *method for Test Problem* 4.

However, preconditioning T_N by K_N cannot improve the ill-conditionedness of matrix T_N. The preconditioned matrix $K_N^{-1}T_N$ is still nearly singular, and the solution is not accurate due to the ill-conditionedness.

For the extreme case $\varepsilon = 0$, matrix T_N is singular with eigenvalues repeated at 0, and T_N^{-1} does not exist. If $\mathbf{b}$ is in the rank space of T_N, $T_N\mathbf{x} = \mathbf{b}$ has more than one solution of the following form,

$$\mathbf{x} = \mathbf{x}^* + \mathbf{v}, \tag{6.1}$$

where $\mathbf{x}^*$ is in the row space of T_N with $T_N\mathbf{x}^* = \mathbf{b}$, and $\mathbf{v}$ is in the null space of T_N. Note that K_N^{-1} exists even though T_N is singular. Since only the inversion of the preconditioner is required in each iteration of preconditioned iterative methods, the iteration can be performed without difficulty. It can be verified that $K_N^{-1}T_N$ has one eigenvalue at 0 and all other $N-1$ eigenvalues repeated exactly at 1. As a consequence of the eigenvalue distribution, the CGS method converges to one solution in form (6.1) after one iteration.

7. Conclusion.

In this paper, we generalized the circulant preconditioning technique from symmetric to nonsymmetric Toeplitz matrices. The resulting preconditioned Toeplitz systems are then solved by various iterative methods such as CGN and CGS. For a large class of Toeplitz matrices, we proved that the singular values of $S_N^{-1}T_N$ and $K_N^{-1}T_N$ are clustered around unity except for a fixed number independent of N. When the generating function is rational, the eigenvalues of $K_N^{-1}T_N$ and $S_N^{-1}T_N$ are classified into clustered eigenvalues and outliers. The number of outliers depends on the order of the rational generating function. The clustered eigenvalues are confined in the disk centered at 1 with the radii $\epsilon_K = O(|t_N| + |t_{-N}|)$ and $\epsilon_S = O(|t_{N-M}| + |t_{1-M}|)$ for $K_N^{-1}T_N$ and $S_N^{-1}T_N$, respectively. Since the eigenvalues of $K_N^{-1}T_N$ are more closely clustered than those of $S_N^{-1}T_N$, preconditioner K_N performs better than S_N for solving rational Toeplitz systems.

Acknowledgments. The authors thank Professor Robert J. Plemmons for helpful discussion. The authors are also grateful to the referees for valuable suggestions, including Test Problem 5 in §6.

REFERENCES

[1] R. H. Chan, *Circulant preconditioners for Hermitian Toeplitz systems*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 542–550.

[2] R. H. Chan and T. F. Chan, *Circulant preconditioners for elliptic problems*, Tech. Rep., Dept. of Mathematics, Univ. of California at Los Angeles, CA, Dec. 1990.

[3] R. H. Chan and X. Jin, *A family of block preconditioners for block systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1218–1235.

[4] R. H. Chan, X. Jin, and M. Yeung, *Circulant preconditioners for complex Toeplitz matrices*, Tech. Rep., Dept. of Mathematics, Univ. of Hong Kong, Apr. 1991.

[5] R. H. Chan and G. Strang, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 104–119.

[6] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

[7] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.

[8] T. Huckle, *Circulant and skew-circulant matrices for solving Toeplitz matrices problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 767–777.

[9] T. K. Ku and C. J. Kuo, *Design and analysis of Toeplitz preconditioners*, IEEE Trans. Signal Processing, 40 (1992), pp. 129–141.

[10] ———, *On the spectrum of a family of preconditioned block Toeplitz matrices*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 948–966.

[11] ———, *A minimum-phase* LU *factorization preconditioner for Toeplitz matrices*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1470–1487.

[12] ———, *Spectral properties of preconditioned rational Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 14 (1993), to appear.

[13] N. M. Nachtigal, S. C. Reddy, and L. N. Trefethen, *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl., 13 (1993), pp. 778–795.

[14] J. G. NAGY AND R. J. PLEMMONS, *Some fast Toeplitz least squares algorithms*, in SPIE Internat. Symp. Optical Applied Sci. Engrg., San Diego, CA, 1991.

[15] Y. SAAD AND M. H. SCHULTZ, GMRES*: A generalized minimum residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[16] P. SONNEVELD, CGS*, a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 36–52.

[17] G. STRANG, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.

[18] ———, *Linear Algebra and Its Applications*, 3rd ed., Harcourt Brace Jovanovich, Inc., Orlando, FL, 1988.

[19] L. N. TREFETHEN, *Approximation theory and numerical linear algebra*, in Algorithms for Approximation II, M. Cox and J. C. Mason, eds., Chapman and Hall, London, 1988.

[20] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, U.K., 1965.

[21] A. YAMASAKI, *New preconditioners based on low-rank elimination*, Tech. Rep. Numerical Analysis 89-10, Dept. of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, Dec. 1989.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 545–552, April 1993

016

ON THE COMPLETION OF PARTIALLY GIVEN TRIANGULAR TOEPLITZ MATRICES TO CONTRACTIONS*

GEIR NÆVDAL[†]

Abstract. In this paper a complete description of the patterns of partial triangular Toeplitz contractions, which are completable to triangular Toeplitz contractions, is given.

Key words. matrix completion, contraction, triangular Toeplitz matrix, partial matrix

AMS(MOS) subject classifications. 47A20, 15A60, 15A57, 47A30

1. Introduction. In this paper we consider the problem of the completion of partially given triangular Toeplitz matrices to contractions. By this we mean the following: Suppose that some of the diagonals of a square triangular Toeplitz matrix are given, and the other ones are to be chosen. Is it possible to choose the elements that are not given in such a way that we obtain a triangular Toeplitz matrix which is a contraction (i.e., with its largest singular value less than or equal to one)?

This is a matrix completion problem. Such problems constitute a current area of research [3]. Among the types of matrix completion problems that have been studied are completion to positive definite matrices, to contractions, and to minimal and maximal rank completions. Positive definite and contractive completion problems are closely related, in fact, all contractive completion problems may be reformulated as positive definite completion problems (see [4]).

For many matrix completion problems there are some obvious conditions that must be satisfied for a completion of a certain class to exist. For contractive completions, all submatrices that only contain given elements must be contractions; for positive definite completions all completely specified principal submatrices must be positive definite. A partially given matrix which satisfies such a condition is called a partial contraction and a partial positive definite matrix, respectively. This terminology will also be used for other properties.

One important area of research has been to determine for which patterns of given elements all partial contractions are completable to contractions, and when a partial positive definite matrix is completable to a positive definite matrix. For the positive definite completion problem, the solution is given by Grone, Johnson, Sá, and Wolkowicz [2], and for completion of partial contractions, by Johnson and Rodman [4]. A partial answer for completion of partial Toeplitz contractions is also given by Johnson and Rodman [5]. In this paper, the solution for completion of partial triangular Toeplitz contractions is given. An interesting open problem in this direction is the partial positive definite Toeplitz completion problem. A conjecture about the result is given in [3].

The problem of completing a partially given triangular Toeplitz matrix to a contraction is closely connected to the problem of Carathéodory–Fejér interpolation with gaps [6]. This problem may be stated as follows.

* Received by the editors March 11, 1991; accepted for publication (in revised form) September 16, 1991.

[†] Division of Mathematical Sciences, The Norwegian Institute of Technology, N–7034 Trondheim, Norway. Present address, Statens sikkerhetshøgskole, Skåregt. 103, N-5500 Haugesund, Norway (gn@ssh.no). This work was done while the author was visiting the Institut Mittag Leffler, Auravägen 17, S-182 62 Djursholm, Sweden.

Let P be a finite subset of the nonnegative integers, with n as the largest element in P. We will call P the pattern of given elements, or just the *pattern*. For $i \in P$ let $c_i \in \boldsymbol{C}$ be given. Does there then exist an $f \in H^\infty(D)$ (the space of bounded analytic functions in the unit disc) such that $\|f\| \leq 1$ and $f(z) = \sum_{i=0}^n \gamma_i z^i + O(z^{n+1})$, where $\gamma_i = c_i$ for $i \in P$?

If $P = \{0, 1, 2, \dots, n\}$, this is the Carathéodory–Fejér interpolation problem, and it is well known that a necessary and sufficient condition for the existence of a solution f with the given properties is that

$$\left\| \begin{pmatrix} c_0 & 0 & \dots & \dots & 0 \\ c_1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ c_n & \dots & \dots & c_1 & c_0 \end{pmatrix} \right\| \leq 1.$$

The theory of the Carathéodory–Fejér interpolation problem is extensively treated in the monograph [1], to which the reader is referred for details and references.

In [6] the problem is solved for $P = \{0, 1, 2, \dots, n-2, n\}$, and it is also shown that for all P's there exists a unique function that minimizes the norm of those which satisfy the interpolation condition.

Motivated by the connection to the Carathéodory–Fejér interpolation problem we number the diagonals for a $(n+1) \times (n+1)$ lower triangular Toeplitz matrix in the following way: The main diagonal is given the number 0 and then the diagonals are numbered in increasing order to the last, which becomes the nth, one.

From the theory of Carathéodory–Fejér interpolation, we know that in the problem we are considering there is no loss of generality in assuming that the last diagonal is given.

The main theorem of this paper is a description of those patterns P with the property that all partial lower triangular Toeplitz contractions are completable to lower triangular Toeplitz contractions. The pattern with this property will be called *completable*. It turns out that the completable patterns are only those of the form $P = \{k + ip : \ k < p, \ 0 \leq i \leq m\}$ for some nonnegative integers k, p, m.

This result depends strongly on the fact that we consider triangular Toeplitz matrices; the completable patterns are different from those obtained in [2], [4], and [5].

2. The main theorem.

THEOREM 2.1. *Let P be a finite subset of the nonnegative integers. Then P is a completable pattern if and only if*

$$P = \{k + ip : k < p, \ i \text{ runs through the integers } 0, 1, \dots, m\}. \tag{2.1}$$

To prove the if part of the theorem is quite easy. For instance, we can look at the corresponding interpolation problem for H^∞ functions, where the coefficients c_j are given for $j \in P$. Let us suppose that

$$\mu = \inf \left\{ \|f\|, f \in H^\infty, f(z) = \sum_{i=0}^n \gamma_i z^i + O(z^{n+1}) \ \text{ where } \ \gamma_j = c_j \ \text{ for } \ j \in P \right\}$$

is attained by a function f_1. Then the function

$$f_2(z) = \frac{f_1(z) + \omega^{-k} f_1(\omega z) + \omega^{-2k} f_1(\omega^2 z) \cdots + \omega^{-(p-1)k} f_1(\omega^{p-1} z)}{p},$$

where $\omega = \exp i\frac{2\pi}{p}$, is also an optimal solution of the interpolation problem. Moreover, this solution may be written as $f_2(z) = z^k g(z^p)$ where $g \in H^\infty$. Since the optimal solution of the interpolation problem above is unique (see [6]), it follows that

$$\mu = \left\| \begin{pmatrix} c_k & 0 & \dots & \dots & 0 \\ c_{k+p} & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ c_n & \dots & \dots & c_{k+p} & c_k \end{pmatrix} \right\|. \tag{2.2}$$

Thus we have proved the first part of the theorem, since the matrix above is a completely given submatrix of the partially given Toeplitz matrix.

It is also possible to prove this part of the theorem by noting that the partially given Toeplitz matrix we are considering is equivalent, by doing row and column operations, to a block matrix where all the elements in the diagonal blocks are given and all the elements outside the blocks on the diagonal are unknowns or zeros. The matrix given in (2.2) is then the matrix on the block diagonal with the largest norm, and by choosing all the unknown elements equal to zero the first part of the theorem is proven.

To prove the only if part we must construct partial triangular Toeplitz contractions that are not completable to triangular Toeplitz contractions. To do this, we use an induction argument, and start working from the lower left part of the matrix.

3. Proof of the only if part of the theorem.

3.1. The first step. We use the following notation when denoting a partially given lower triangular Toeplitz matrix, T. Let c_i denote given elements, x_i elements which are not given, and y_i elements placed in such a position that we want both possibilities to be considered. To have a short notation for a partially given lower triangular Toeplitz matrix, we define

$$T(\gamma_0, \gamma_1, \dots, \gamma_n) = \begin{pmatrix} \gamma_0 & 0 & \dots & \dots & 0 \\ \gamma_1 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \gamma_n & \dots & \dots & \gamma_1 & \gamma_0 \end{pmatrix}.$$

Since in the rest of the paper we discuss only lower triangular Toeplitz matrices, we will henceforth denote a lower triangular Toeplitz contraction by the term "contraction," and we will denote a partial lower triangular Toeplitz contraction by the term "partial contraction."

The first three lemmas show that the theorem holds with the extra assumption that $n-1, n \in P$.

LEMMA 3.1. *Let* $c_n = c_{n-1} = (1/\sqrt{2})$ *and* $c_i = 0$ *if* $i \neq n-1, n$. *Then*

$$T(y_0, \ldots, y_{n-3}, x_{n-2}, c_{n-1}, c_n)$$

is a partial contraction, which cannot be completed into a contraction.

Proof. This follows since

$$\left\| \begin{pmatrix} \frac{1}{\sqrt{2}} & x \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \right\| = 1$$

if and only if $x = -(1/\sqrt{2})$. □

LEMMA 3.2. *If* $c_n = 1/10, c_{n-1} = 1/4, c_{n-2} = 737/792$, *and* $c_i = 0$ *for* $i \neq n-2, n-1, n$, *then*

$$T = T(y_0, \ldots, y_{n-4}, x_{n-3}, c_{n-2}, c_{n-1}, c_n)$$

is a partial contraction, which is not completable to a contraction.

Proof. First we must check that all the submatrices which contain the elements c_{n-2}, c_{n-1}, c_n are contractions. The elements c_{n-2}, c_{n-1}, c_n are chosen in such a way that

$$\left\| \begin{pmatrix} c_{n-1} & c_{n-2} & x_{n-3} \\ c_n & c_{n-1} & c_{n-2} \end{pmatrix} \right\| = 1$$

for a unique x_{n-3}. But for this x_{n-3},

$$\left\| \begin{pmatrix} c_n & c_{n-1} & c_{n-2} & x_{n-3} \end{pmatrix} \right\| > 1,$$

which completes the proof. □

LEMMA 3.3. *Let* $p > 3$, $c_n = c_{n-(p/2)} = (1/\sqrt{2})$ *if* p *is even,* $c_{n-1} = c_{n-((p+1)/2)} = (1/\sqrt{2})$ *if* p *is odd; and let* $c_i = 0$ *for* $i \neq n - \frac{p}{2}, n$ *if* p *is even, and* $i \neq n - \frac{p+1}{2}, n-1$ *if* p *is odd. Then*

$$T = T(y_0, \ldots, y_{n-p-1}, x_{n-p}, c_{n-p+1}, \ldots, c_n)$$

is a partial contraction, which is not completable to a contraction.

Proof. If p is even note that

$$\left\| \begin{pmatrix} c_{n-(p/2)} & x_{n-p} \\ c_n & c_{n-(p/2)} \end{pmatrix} \right\| = 1$$

if and only if $x_{n-p} = -(1/\sqrt{2})$. If p is odd, the argument is similar. □

Let us now consider the case

$$T = T(y_0, \ldots, y_{k-1}, c_k, \ldots, c_p, x_{p+1}, \ldots, x_{n-1}, c_n).$$

LEMMA 3.4. *If* $c_i = 0$ *for* $i \leq p-2$,

$$\left\| \begin{pmatrix} c_p & c_{p-1} \end{pmatrix} \right\| < 1,$$

and

$$\left\| \begin{pmatrix} c_{p-1} & 0 \\ c_p & c_{p-1} \end{pmatrix} \right\| = 1,$$

then there exist several choices of c_n such that

$$T = T(y_0, \dots, y_{p-3}, c_{p-2}, c_{p-1}, c_p, x_{p+1}, \dots, x_{n-1}, c_n)$$

is a partial contraction. But there is only one choice of c_n such that T is completable to a contraction.

Proof. Since it is possible to choose x_i, $i < n$ such that

$$T' = T(y_0, \dots, y_{p-3}, c_{p-2}, c_{p-1}, c_p, x_{p+1}, \dots, x_{n-1})$$

is a contraction, all the completely given submatrices of T' must be contractions.

All the completely given submatrices of T that contain c_n are submatrices of a matrix of the form

$$A = \begin{pmatrix} 0 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & & & & \vdots \\ 0 & \vdots & & & & \vdots \\ c_{p-1} & \vdots & & & & \vdots \\ c_p & 0 & \dots & \dots & \dots & 0 \\ c_n & c_p & c_{p-1} & 0 & \dots & 0 \end{pmatrix},$$

since $c_i = 0$ for $i < p$.

For every completely given submatrix there exists a disc with center at the origin such that if c_n is placed inside this disc the submatrix becomes a contraction. (This follows by [1, Thm. IV.3.1].) The possible choices of c_n in the lemma are then given by the smallest of these discs.

From the theory of the Carathéodory–Fejér interpolation problem it is well known that there exists a unique c_n such that T is a contraction. □

We also need a counterexample for the case

$$T = T(y_0, \dots, y_{p-3}, x_{p-2}, c_{p-1}, c_p, x_{p+1}, \dots, x_{n-1}, c_n).$$

LEMMA 3.5. *For $i < p-2$ let $c_i = 0$, $c_{p-1} = c_p = (1/\sqrt{2})$ and $c_n = 0$. Then*

$$T = T(y_0, \dots, y_{p-3}, x_{p-2}, c_{p-1}, c_p, x_{p+1}, \dots, x_{n-1}, c_n)$$

is a partial contraction, which cannot be completed to a contraction.

Proof. Here the situation is described by looking at

$$\begin{pmatrix} \ddots & & & & \\ \ddots & x_{p-2} & & & \\ \ddots & \frac{1}{\sqrt{2}} & x_{p-2} & & \\ \ddots & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & x_{p-2} & \\ & x_{p+1} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \ddots \\ & & \ddots & \ddots & \ddots \end{pmatrix},$$

and it easily follows that all the completely given submatrices must be contractions. However, it is also well known that

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & x_{p-2} \\ x_{p+1} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

is not completable to a contraction [5]. □

The main ideas for constructing the necessary examples are now introduced. We show that these ideas can be used to construct all the examples we need by doing an induction step.

3.2. The induction step. Before continuing to the induction step, let us introduce the following notation for describing the pattern of a partially given lower triangular Toeplitz matrix. Let C_r denote that we have r consecutive diagonals with given elements, let X_s denote that we have s consecutive diagonals with unknown elements, and let Y_j denote that j consecutive diagonals are given where we want to consider every possible structure. Then, for instance, we have

$$\begin{aligned} T &= T(y_0, \dots, y_j, c_{j+1}, \dots, c_{j+t}, x_{j+t+1}, \dots, x_{j+t+s}, c_{j+t+s+1}, \dots, c_{j+t+s+r}) \\ &= T(Y_{j+1}, C_t, X_s, C_r). \end{aligned}$$

Suppose that we have proved that the only possibilities for all partial contractions to be completable to a contraction are when

$$T = T(Y_k, C_1, X_{p-1}, C_1, X_{p-1}, C_1, X_{p-1}, C_1, \dots, \dots, X_{p-1}, C_1, X_{p-1}, C_1).$$

Then the theorem follows if we can prove that T, in fact, must be on one of the forms

$$T = T(X_k, C_1, X_{p-1}, C_1, X_{p-1}, C_1, X_{p-1}, C_1, \dots, \dots, X_{p-1}, C_1, X_{p-1}, C_1)$$

with $k < p$ or

$$T = T(Y_{k-p}, C_1, X_{p-1}, C_1, X_{p-1}, C_1, X_{p-1}, C_1, \dots, \dots, X_{p-1}, C_1, X_{p-1}, C_1).$$

Let us construct the partial contractions, which are not completable to contractions for the other possibilities. There are four other possibilities that we need to consider:

$$\begin{aligned} T = T(&Y_{k-p-1}, C_2, X_{p-1}, C_1, X_{p-1}, C_1, X_{p-1}, C_1, \dots, \\ &\dots, X_{p-1}, C_1, X_{p-1}, C_1) \end{aligned} \tag{3.1}$$

and

$$T = T(Y_{k-s-1}, C_1, X_s, C_1, X_{p-1}, C_1, X_{p-1}, C_1, \dots, \dots, X_{p-1}, C_1, X_{p-1}, C_1) \tag{3.2}$$

with $s < p-1$ or $s > p-1$, and

$$T = T(X_k, C_1, X_{p-1}, C_1, X_{p-1}, C_1, \dots, \dots, X_{p-1}, C_1, X_{p-1}, C_1) \tag{3.3}$$

with $k \geq p$. (Of course, we assume that $p \geq 2$.)

Let us start with the case given by (3.1). Consider the possibilities of w in

$$T = \begin{pmatrix} c_{k-p-1} & \cdots & \cdots & w \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ c_{n-p} & \cdots & \cdots & c_{k-p-1} \end{pmatrix}. \tag{3.4}$$

If the element w either belongs to the upper triangular part of T or is a given element, let $c_i = 0$ if $i \neq k-p-1, n-p, n$; let

$$\left\| \begin{pmatrix} c_{n-p} & c_{k-p-1} \end{pmatrix} \right\| < 1;$$

and let

$$\left\| \begin{pmatrix} c_{k-p-1} & 0 \\ c_{n-p} & c_{k-p-1} \end{pmatrix} \right\| = 1.$$

Then use the same kind of argument as in Lemma 3.4. If $w = x_j$ for some j, let $c_{n-p} = c_{k-p-1} = (1/\sqrt{2})$ and $c_i = 0$ for $i \neq k-p-1, n-p$, and use the same kind of argument as in Lemma 3.3.

The case given by (3.2) with $s < p-1$ follows by a similar argument as the one above, but now with c_{k-s} instead of c_{k-p-1} in (3.4) (and Lemma 3.5 instead of Lemma 3.3).

The theorem follows by proving the cases given by (3.2) with $s > p-1$, and by (3.3). These two cases are now completed by showing that the ideas of Lemmas 3.1–3.3 may also be used here.

By doing row and column operations we may transform the matrix to a block matrix of the form

$$\begin{pmatrix} A_0 & B_{1\,2} & \cdots & \cdots & \cdots & B_{1\,p+1} \\ B_{2\,1} & A_1 & \ddots & & & \vdots \\ \vdots & \ddots & A_1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & A_1 & B_{p\,p+1} \\ B_{p+1\,1} & \cdots & \cdots & \cdots & B_{p+2\,p+1} & C \end{pmatrix},$$

where

$$A_0 = \begin{pmatrix} y_{n-mp} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & & \vdots \\ y_{n-(k+1)p} & & \ddots & \ddots & & & \vdots \\ x_{n-kp} & \ddots & & \ddots & \ddots & & \vdots \\ c_{n-(k-1)p} & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & 0 \\ c_n & \cdots & c_{n-(k-1)p} & x_{n-kp} & y_{n-(k+1)p} & \cdots & y_{n-mp} \end{pmatrix},$$

$$A_1 = \begin{pmatrix} y_{n-mp} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & & \vdots \\ y_{n-(k+1)p} & & \ddots & \ddots & & & \vdots \\ x_{n-kp} & \ddots & & \ddots & \ddots & & \vdots \\ c_{n-(k-1)p} & \ddots & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & 0 \\ c_{n-p} & \cdots & c_{n-(k-1)p} & x_{n-kp} & y_{n-(k+1)p} & \cdots & y_{n-mp} \end{pmatrix},$$

and the blocks outside the block diagonal and the block C consist of zeros and elements that are not specified. (The block C may disappear, together with the corresponding row and column in the block matrix.) The last step in the proof of the theorem now follows by using the constructions in Lemmas 3.1–3.3 to construct partial contractions, which are not completable to contractions.

Acknowledgments. The author wants to thank the Mittag–Leffler Institute for its hospitality. He also wants to thank Professor H. H. Martens and the referees for some useful comments.

REFERENCES

[1] C. Foias and A. E. Frazho, *The Commutant Lifting Approach to Interpolation Problems*, Operator Theory: Advances and Applications, Vol. 44, Birkhäuser-Verlag, Basel, 1990.

[2] R. Grone, C. R. Johnson, E. Sá, and H. Wolkowicz, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 24 (1984), pp. 109–124.

[3] C. R. Johnson, *Matrix Completion Problems: A Survey*, in Matrix Theory and Applications, Proceedings of Symposia in Applied Mathematics, Vol. 40, C. R. Johnson, ed., American Mathematical Society, Providence, RI, 1990.

[4] C. R. Johnson and L. Rodman, *Completion of partial matrices to contractions*, J. Funct. Anal., 69 (1986), pp. 260–267.

[5] ———, *Completion of Toeplitz partial contractions*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 159–167.

[6] G. Nævdal, *On Carathéodory–Fejér Interpolation with Gaps*, Preprint, Rep. No. 3, Institut Mittag Leffler, Djursholm, Sweden, 1990/1991.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 553–559, April 1993

017

BUNCH–KAUFMAN FACTORIZATION FOR REAL SYMMETRIC INDEFINITE BANDED MATRICES*

MARK T. JONES[†] [‡] AND MERRELL L. PATRICK[†]

Abstract. The Bunch–Kaufman algorithm for factoring symmetric indefinite matrices has been rejected for banded matrices because it destroys the banded structure of the matrix. Herein, it is shown that for a subclass of real symmetric matrices which arise in solving the generalized eigenvalue problem using the Lanczos method, the Bunch–Kaufman algorithm does not result in major destruction of the bandwidth. Space/time complexities of the algorithm are given and used to show that the Bunch–Kaufman algorithm is a significant improvement over banded LU factorization. Timing comparisons are used to show the advantage held by the authors' implementation of Bunch–Kaufman over the implementation of the multifrontal algorithm for indefinite factorization in MA27 when factoring this subclass of matrices.

Key words. Bunch–Kaufman factorization algorithm, symmetric indefinite matrices, banded matrices

AMS(MOS) subject classifications. 65F05, 65F15

1. Introduction.

The Bunch–Kaufman algorithm is considered one of the best methods for factoring full, symmetric indefinite matrices [3], [1]. For sparse matrices, multifrontal methods that use pivoting techniques similar to the Bunch–Kaufman algorithm have been developed [4]. Bunch and Kaufman [3] have shown that for tridiagonal and pentadiagonal matrices the Bunch–Kaufman algorithm does not suffer from catastrophic fill-in. To date, for matrices with a semibandwidth[1] larger than 3, the Bunch–Kaufman algorithm has been thought to destroy the banded structure of the matrix [3]; therefore, either banded LU factorization or the multifrontal variant of the Bunch–Kaufman algorithm had to be used.

Herein it is shown that for a subclass of real symmetric indefinite matrices, which arise in solving the generalized eigenvalue problem using Lanczos' method, the Bunch–Kaufman algorithm does not result in major destruction of the bandwidth. Furthermore, for this class of problems, the Bunch–Kaufman factorization algorithm is a significant improvement over LU factorization and the multifrontal method. In addition to taking advantage of symmetry, the Bunch–Kaufman algorithm yields the inertia of the matrix essentially for free[2] [3], which is important in eigenvalue calculations. LU factorization does not yield the inertia as a by-product and destroys the symmetry of the matrix, thus increasing storage requirements for its implementation.

In §2 we give one of the several variations of the Bunch–Kaufman algorithm and in §3 describe a subclass of matrices to which we apply it. An efficient implementation of the method is described in §4. The space/time complexity of the implementation is discussed in §5. Comparisons to other algorithms are given in §6. Conclusions are drawn in §7.

* Received by the editors April 9, 1990; accepted for publication (in revised form) November 14, 1991. This research was supported by National Aeronautics and Space Administration contracts NAS1-18107 and NAS1-18605 and by Air Force Office of Scientific Research grant 88-0117 while the authors were in residence at the Institute for Computer Applications in Science and Engineering. Additional support was provided by NASA grant NAG-1-466.

[†] Department of Computer Science, Duke University, Durham, North Carolina 27706.

[‡] Present address, Argonne National Laboratory, MCS Division, 9700 South Cass Ave, Argonne, Illinois 60439 (mjones@mcs.anl.gov).

[1] We include the diagonal in our definition of semibandwidth.

[2] The multifrontal variant of Bunch–Kaufman also has these advantages.

```
1)  i = 1
2)  while ($i \le n$)
3)        $\lambda = \max_{j=i+1,n} \mid a_{j,i} \mid$
4)        set $r$ to the row number of this value
5)        if $\lambda\alpha < \mid a_{i,i} \mid$ then
6)          perform a $1 \times 1$ pivot
7)          i = i + 1
8)        else
9)          $\sigma = \max_{j=i+1,n} \mid a_{r,j} \mid$
10)         if $\alpha\lambda^2 < \sigma \mid a_{i,i} \mid$ then
11)           perform a $1 \times 1$ pivot
12)           i = i + 1
13)         else
14)            exchange rows and columns $r$ and $i + 1$
15)            perform a $2 \times 2$ pivot
16)            i = i + 2
17)         endif
18)       endif
19) endwhile
20) if inertia is desired, then scan the $D$ matrix
```

FIG. 1. *The Bunch–Kaufman factorization algorithm.*

2. The Bunch–Kaufman algorithm.

The Bunch–Kaufman algorithm factors A, an $n \times n$ real symmetric indefinite matrix, into LDL^T while doing symmetric permutations on A to maintain stability, resulting in the following equation:

$$PAP^T = LDL^T, \tag{1}$$

where P is a permutation matrix, L is lower triangular, and D is a block diagonal matrix consisting of 1×1 and 2×2 blocks. Because A is indefinite, the usual LDL^T decomposition where D is diagonal may not exist. The Cholesky decomposition LL^T fails because it attempts to take the square root of a negative number when A is indefinite. Although several variations of the algorithm exist, algorithm D of the Bunch–Kaufman paper is the least destructive of the banded structure [3]. The algorithm is shown in Fig. 1.

The parameter in steps 5 and 10, α, is chosen such that stability is maximized and has been shown by Bunch and Kaufman to be approximately 0.525 [3]. The exchange of rows and columns in step 13 maintains the symmetry of the matrix, unlike LU factorization which destroys the symmetry of the matrix by permuting only rows.

3. Applicable set of matrices.

Bunch and Kaufman show that, in general, if m is the semibandwidth of a matrix being factored, then a 2×2 pivot can increase the semibandwidth of the remaining nonzeros from m to $(2m)-2$ and that this can happen every other step, thus resulting in the complete destruction of the band structure due to fill-in outside the band [3]. However, it is shown in §4 that for a subclass of matrices this fill-in can be controlled. The number of 2×2 pivots is bounded by the number of negative eigenvalues of A because each 2×2 pivot represents a positive-negative eigenvalue pair [3]. Also, the increase of the semibandwidth from m to $(2m) - 2$ is a worst case that is not likely to occur in practice. Therefore, for matrices with a small number of negative eigenvalues (in relation to the size of the matrix), it is possible to use Bunch–Kaufman factorization with very little fill-in. Such matrices arise in eigenvalue calculations where the smallest eigenvalues are sought. Methods such as inverse iteration and Lanczos's method are often used to find the smallest eigenvalues of a matrix A. These methods often require the factorization of a matrix $(A - \sigma I)$, where σ is normally very near the left end of A's spectrum, but may not be to the left of the smallest eigenvalue. Thus the matrix is indefinite [5] but has only a small number of negative eigenvalues. These matrices can be banded, as they often are in structural mechanics [2]. The difficulty is that the location and amount of the fill-in outside the band is not possible to predict a priori. In the following section, an algorithm that dynamically allows for fill-in during factorization is presented.

```
.   .                                           .   .
.   .                                           .   .
● ● ● ● ● x                      ← row i →      ● ● ● ● ● x
0 ● ● ● ● x x                  ← row i+1 →      0 ● ● ● ● x x
0 0 ● ● ● x x x                                 0 0 ● ● ● x x x
0 0 0 ● ● x x x x                               0 0 0 ● ● x x x x
0 0 0 0 ● x x x x x                             0 0 0 0 ● x x x x x
0 0 0 0 0 x x x x x x            ← row r →      0 0 0 0 0 x x x x x x
0 0 0 0 0 0 x x x x x x                         0 0 0 0 0 f x x x x x x
0 0 0 0 0 0 0 x x x x x x                       0 0 0 0 0 f f x x x x x x
0 0 0 0 0 0 0 0 x x x x x x                     0 0 0 0 0 f f f x x x x x x
0 0 0 0 0 0 0 0 0 x x x x x x                   0 0 0 0 0 f f f f x x x x x x
0 0 0 0 0 0 0 0 0 0 x x x x x x                 0 0 0 0 0 f f f f f x x x x x x
0 0 0 0 0 0 0 0 0 0 0 x x x x x x               0 0 0 0 0 0 0 0 0 0 0 x x x x x x
.   .   .                                       .   .   .
.   .   .                                       .   .   .
```

FIG. 2. *Example of fill-in (note: this is an example of worst-case fill-in).*

4. Efficient implementation of the algorithm.

As the algorithm in Fig. 3 is executed, the original matrix is copied, piecewise, from one place in memory to another. This allows for dynamic allocation of fill-in as well as only requiring part of the matrix to be in main memory at any particular time. Fill-in only takes place in a small triangle when a 2×2 pivot occurs. If a pivot occurs at step i, this triangle is of the form shown in Fig. 2, where •'s represent eliminated elements in L, x's are uneliminated nonzeros, 0's are zeros outside the band for which no storage is needed, and f's are areas where fill occurs. The triangle of fill is from row $r+1$ to row $r+m$, where m is the semibandwidth (this area may already contain nonzeros depending on the value of r, so no extra memory may be needed). The algorithm is given in Fig. 3. P is a vector representing the permutation matrices. The only time fill outside the band occurs in step 14 of the algorithm when a 2×2 pivot occurs and then storage for the fill is allocated dynamically.

5. Speed and storage analysis.

In this section the space/time requirements of this implementation of the Bunch–Kaufman algorithm are compared with those of banded LU factorization. The storage requirements for both algorithms are analyzed for two different situations: (1) when factoring a matrix that falls in the subclass described in §2; and (2) when factoring a matrix pencil such as $(K - \sigma M)$, where K and M are symmetric positive definite and σ is near the left end of the eigenspectrum of $Kx = \lambda Mx$.

In the first situation, the storage required by the algorithm presented in §4 is significantly less than that required by LU factorization for the set of matrices that was described in §2. The storage requirements of LU depend on the pivot selections made with the best-case storage requirement being $2mn$ and the worst-case storage requirement being $3mn$. The storage needed by banded Bunch–Kaufman is mn for the original storage from which the matrix is copied, plus mn for the locations to which the matrix is copied, plus an additional amount C, which is the amount of

```
1)  $upto = 0$
2)  $i = 0$
3)  While $(i \leq n)$
4)          read rows $upto + 1$ to $\min(n,i+m)$ of the matrix $A$ into $L$
C             (no extra space for fill needs to be added for these rows)
5)          $upto = \min(n,i+m,upto)$
6)          $\lambda = \max_{j=i+1,upto} \mid a_{j,i} \mid$
7)          set $r$ to the row number of $\lambda$
8)          if $\lambda\alpha \leq\mid a_{i,i} \mid$ then goto 11
9)          $\sigma = \max_{j=i+1,upto} \mid a_{r,j} \mid$
C             (it may be necessary to access some elements that are not
C             read in at this point, but the number of elements is small,
C             so they may be read into $L$ or simply discarded,
C             this is only a concern if i/o is taking place)
10)         if $\alpha\lambda^2 \leq \sigma \mid a_{i,i} \mid$ then
11)           perform a $1 \times 1$ pivot
12)           $i = i + 1$
13)         else
C                permute the matrix
14)              read rows $upto + 1$ to $\min(n,r+m)$ of the matrix $A$ into $L$,
                 and allocate space for the fill triangle
15)              $upto = \min(n,r+m,upto)$
16)              exchange rows and columns $r$ and $i+1$
17)              $p_i = i$
18)              $p_{i+1} = r$
19)              perform a $2 \times 2$ pivot
20)              $i = i + 2$
21)         endif
22) endwhile
```

FIG. 3. *The banded Bunch–Kaufman factorization algorithm.*

storage necessary for the fill-in triangles, giving a total storage of $2mn + C$. Because of the small number of negative eigenvalues, C is much less than mn. The D matrix requires no additional storage: The diagonal of D can be stored in the diagonal of L[3] and the subdiagonal of D can be stored in L because if a 2×2 pivot occurs at step i, then $l_{i,i+1}$ is 0 and therefore need not be stored. When C is small, approximately mn storage locations are saved by using the Bunch–Kaufman algorithm instead of LU factorization.

In the second situation (which arises in an efficient implementation of Lanczos's method for solving $Kx = \lambda Mx$), the shift σ may change during execution of the algorithm, so K and M must be saved throughout the computation. In this situation, the storage requirements[4] for LU factorization are increased by $2mn$, but the storage needed by Bunch–Kaufman increases by mn, making it even more attractive in this case.

The operation counts for situations 1 and 2 are the same, however, the operation count for Bunch–Kaufman is significantly less than that of LU factorization because symmetry is exploited and the fill-in is limited. For simplicity, the operations added by the fill-in during Bunch–Kaufman are ignored, since this amount is very small. The high-order term in the operation counts for Bunch–Kaufman is approximately nm^2 arithmetic operations plus approximately $3nm$ comparisons[5], while the high-order

[3] The diagonals of L are all 1.

[4] These bounds include the space for K and M.

[5] This is an upper bound, calculated by adding the cost of the search for λ, nm, to the worst-case cost of the search for σ, $2nm$. Because the search for σ usually does not occur at every step and does not usually cost $2m$, this bound is very pessimistic.

TABLE 1
Operation counts for factorization: $n = 1824$, $m = 240$.

Method	No. of neg. eigenvalues	No. of 2 × 2 pivots	Adds.	Mults.	Comps.	Fill
Chol.	N/A	N/A	44433080	48140336	0	0
B–K	5	4	48277445	48686784	446326	2083
LU	5	N/A	137241687	137648943	409079	658463
B–K	19	7	52023663	52445756	485452	14837
LU	19	N/A	137412094	137819350	409079	659837

term for the operation counts for LU is approximately $4nm^2$ arithmetic operations plus approximately nm comparisons.

6. Comparisons with other methods. In this section, the implementation of banded Bunch–Kaufman is compared with two other classes of methods: methods using banded data structures and methods using sparse data structures.

In the first set of comparisons, banded Bunch–Kaufman is compared to banded Cholesky factorization and banded LU factorization. The comparison to Cholesky factorization is given only as a reference point; the matrices are shifted before Cholesky factorization to make them positive definite to allow the algorithm to succeed. The *sgbfa* subroutine from LINPACK is used for the banded LU factorization. Because methods using band data structures can be implemented in a similar fashion and rely on the same computer operations, operation counts are used for the comparison rather than execution times. The algorithms are compared using four matrices: the first two matrices have a relatively wide bandwidth with 5 and 19 negative eigenvalues, and the second two matrices have a small bandwidth with 5 and 15 negative eigenvalues.

Examination of the results in Tables 1 and 2 shows Bunch–Kaufman to be slightly more expensive than Cholesky factorization in terms of add and multiply operations. Banded LU factorization is shown to be significantly more expensive than Bunch–Kaufman in terms of add and multiply operations and only slightly less expensive in terms of comparisons. The number of operations required by LU factorization falls short of the bound given in the previous section because the worst-case choice of pivots is not made by LU and therefore the number of operations performed is reduced. LU required significantly more fill-in than the Bunch–Kaufman algorithm.[6] The results also show that as the number of negative eigenvalues increases, the fill-in for Bunch–Kaufman tends to increase. Part of the reason for the small fill-in required by Bunch–Kaufman is that the number of 2 × 2 pivots required tends to fall short of the number of negative eigenvalues.

In the second set of comparisons, banded Bunch–Kaufman is compared to the multifrontal method for indefinite sparse matrices given by Duff and Reid [4]. The implementation of the multifrontal algorithm, MA27, from the Harwell library is used. This set of comparisons is designed to show that for the class of banded matrices in this paper, the banded Bunch–Kaufman algorithm is faster than the multifrontal algorithm. Because the operations and data structures in the two algorithms are significantly different, execution times from a Sun/4 are used as the basis for comparison.

[6] Although *sgbfa* requires $2mn$ fill-in, it could be implemented in a similar fashion to the banded Bunch–Kaufman algorithm and, therefore, take as little as mn fill-in. For this reason, the actual fill-in outside the band, including the mn fill-in due to loss of symmetry, is reported in Tables 1 and 2.

TABLE 2
Operation counts for factorization: $n = 1980$, $m = 59$.

Method	No. of neg. eigenvalues	No. of 2 × 2 pivots	Adds.	Mults.	Comps.	Fill
Chol.	N/A	N/A	3321051	3434180	0	0
B-K	5	3	3322513	3435664	142978	22
LU	5	N/A	10342067	10455196	115108	175206
B-K	15	3	3324670	3437856	152370	57
LU	15	N/A	10618321	10731450	115108	178583

TABLE 3
Comparison of banded Bunch–Kaufman and MA27.

Method	N	M	No. of neg. eigenvalues	Time in seconds
B-K	1809	253	5	163.59
MA27	1809	253	5	516.60
B-K	1809	253	19	177.53
MA27	1809	253	19	541.55
B-K	1980	59	5	10.44
MA27	1980	59	5	38.78
B-K	1980	59	15	12.47
MA27	1980	59	15	52.68

Matrices similar to those in the first comparison are used. In this comparison, MA27 is directed to use the natural banded ordering because: (1) factorization in MA27 is faster when the natural banded ordering is used than when MA27 uses its own ordering; and (2) MA27 achieves an additional savings in time when it does not have to produce its own ordering. An examination of the results in Table 3 shows that banded Bunch–Kaufman holds a significant advantage over the multifrontal method. The major reason for this advantage is the large overhead incurred when using sparse data structures. The only conclusion that should be drawn from this comparison is that for this class of matrices, the banded Bunch–Kaufman algorithm is preferable to the multifrontal method as implemented in MA27.

7. Conclusions. An algorithm based on the Bunch–Kaufman method for factoring banded symmetric indefinite matrices has been presented that greatly limits fill-in outside the band and takes advantage of the symmetry of the matrix. This method was shown to be a more efficient factorization method than LU factorization in terms of time and storage for factoring symmetric indefinite matrices with a small number of eigenvalues. Banded Bunch–Kaufman was also shown to be faster than an implementation of the multifrontal method for this same class of problems. Bunch and Kaufman have previously shown this method to be nearly as stable as LU factorization [3].

REFERENCES

[1] V. BARWELL AND A. GEORGE, *A comparison of algorithms for solving symmetric indefinite systems of linear equations*, ACM Trans. Math. Software, 2 (1976), pp. 242–251.

[2] C. P. BLANKENSHIP AND R. J. HAYDUK, *Potential supercomputer needs for structural analysis*, Presentation at the Second Internat. Conf. Supercomputing, May 3–8, 1987.

[3] J. R. BUNCH AND L. KAUFMAN, *Some stable methods for calculating inertia and solving symmetric linear systems*, Math. Comput., 31 (1977), pp. 163–179.

[4] I. Duff and J. Reid, *The multifrontal solution of indefinite sparse symmetric linear equations*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.

[5] B. Nour-Omid, B. N. Parlett, and R. L. Taylor, *Lanczos versus subspace iteration for solution of eigenvalue problems*, Internat. J. Numer. Methods Engrg., 19 (1983), pp. 859–871.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 560–587, April 1993

NEWTON METHODS FOR LARGE-SCALE LINEAR EQUALITY-CONSTRAINED MINIMIZATION*

A. FORSGREN[†] AND W. MURRAY[‡]

Abstract. Newton methods for large-scale minimization subject to linear equality constraints are discussed. For large-scale problems, it may be prohibitively expensive to reduce the problem to an unconstrained problem in the null space of the constraint matrix. We investigate computational schemes that enable the computation of descent directions and directions of negative curvature without the need to know the null-space matrix. The schemes are based on factorizing a sparse symmetric indefinite matrix. Three different methods are proposed based on the schemes described for computing the search directions. Convergence properties for the methods are established.

Key words. linear equality-constrained minimization, negative curvature, modified Newton method, symmetric indefinite factorization, large-scale minimization, linesearch method

AMS(MOS) subject classifications. 49D37, 65K05, 90C30

1. Introduction.

We consider methods for finding a local minimizer of the problem

$$\begin{array}{ll} \underset{x\in\Re^n}{\text{minimize}} & f(x) \\ \text{subject to} & Ax = b, \end{array} \tag{1.1}$$

where A is an $m \times n$ matrix and $f \in C^2$. We are interested in the case when n and possibly m are large and when second derivatives of f are available. It is assumed that A is a sparse matrix of full row rank. We also assume that an initial feasible point x_0 is known, and that the level set $S(x_0) = \{x : f(x) \le f(x_0),\ Ax = b\}$ is compact.

Problem (1.1) only involves linear equality constraints. Our main reason for considering such problems is to focus on certain computational and theoretical issues that arise when solving large-scale constrained problems. If linear inequality constraints are present, or some constraint functions are nonlinear, further difficulties arise. However, even for these cases, problems of the type (1.1) may arise as subproblems; for example, if the inequality constraints are removed by a barrier transformation, and the nonlinear constraints are linearized.

If (1.1) is solved by Newton's method, the step p from the current iterate x may be defined as $p = Zp_Z$, where

$$Z^THZp_Z = -Z^Tg. \tag{1.2}$$

The columns of the matrix Z form a basis for the null space of A, H denotes the Hessian matrix $\nabla^2 f(x)$, and g denotes the gradient $\nabla f(x)$. Throughout the paper, Z is used to denote *any* fixed matrix whose columns form a basis for the null space of A. We shall refer to the matrix Z^THZ as the *reduced Hessian* and the vector Z^Tg

* Received by the editors July 9, 1990; accepted for publication (in revised form) October 30, 1991. This research was partially supported by National Science Foundation grant ECS-8715153, the Office of Naval Research grant N00014-90-J-1242, the Swedish National Board for Technical Development, and the Göran Gustafsson Foundation.

[†] Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden.

[‡] Department of Operations Research, Stanford University, Stanford, California 94305 (walter@sol-walter.stanford.edu).

as the *reduced gradient.* A mathematically equivalent way to obtain p is to solve the equation

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ -\mu \end{pmatrix} = \begin{pmatrix} -g \\ 0 \end{pmatrix}. \tag{1.3}$$

We shall refer to the matrix on the left-hand side of (1.3) as the *KKT-matrix* and denote it by K.

Given p, the next iterate is obtained as $x + p$. If Newton's method converges to a solution of (1.1) and the reduced Hessian is positive definite and Lipschitz continuous, convergence occurs at a quadratic rate. However, Newton's method may not converge from every starting point, and in the neighborhood of a saddle-point or a local maximizer, a sequence of iterates generated by Newton's method may converge to such a point. Consequently, if a method that generates a sequence of improving estimates is required, some modification is needed. If Z^THZ is positive definite, p is a *feasible descent direction*, i.e., a direction p such that $p^Tg < 0$ and $Ap = 0$. In this case, the objective function value at $x+p$ may not be reduced, but a step along p may be found that yields the next improved iterate. If Z^THZ is not positive definite, the Newton direction may not be a direction along which the objective function decreases.

Modifications of Newton's method suitable for iterates where the Newton step does not yield a *sufficient* decrease of the objective function are known. The methods can be divided into two types, *trust-region methods* and *linesearch methods.* See Moré and Sorensen [MS84] and Dennis and Schnabel [DS89] for discussions of these different types of method. We focus on linesearch methods in this paper and in order to simplify the notation, we shall refer to modifications of Newton's method of linesearch type as *modified Newton methods.* Such methods involve the computation of a feasible descent direction s and, if the reduced Hessian has at least one negative eigenvalue, a *feasible direction of negative curvature d.* A direction d is said to be a feasible direction of negative curvature if $d^THd < 0$ and $Ad = 0$. If s is a direction of *sufficient* descent and d is a direction of *sufficient* negative curvature, the convergence of a modified Newton method for linear equality-constrained problems follows from known results on modified Newton methods; see, for example, Fiacco and McCormick [FM68], Gill and Murray [GM74], McCormick [McC77], Fletcher and Freeman [FF77], Mukai and Polak [MP78], Kaniel and Dax [KD79], Moré and Sorensen [MS79], Goldfarb [Gol80] and Forsgren, Gill, and Murray [FGM89b]. We require the computed directions s and d to be feasible, i.e., $As = 0$ and $Ad = 0$. If the reduced Hessian is known and has at least one negative eigenvalue, a direction $\bar{d}$ such that $\bar{d}^TZ^THZ\bar{d} < 0$ may be obtained using techniques for unconstrained problems. Consequently, $d = Z\bar{d}$ is a feasible direction of negative curvature. Similarly, a positive definite modification of Z^THZ enables the computation of a feasible descent direction. Feasibility is therefore not an issue if the reduced Hessian is known, whereas it is not immediately apparent how to satisfy $As = 0$ and $Ad = 0$ when utilizing K.

Although (1.2) and (1.3) are mathematically equivalent, they differ in the amount of work required to obtain the search direction p. To obtain p from (1.2), the reduced Hessian is required. A matrix Z may be obtained by forming the LU-factorization of A; see Gill, Murray, Saunders, and Wright [GMSW87a]. Even if Z is sparse, the matrix Z^THZ may be completely dense, and the amount of work required to form Z^THZ explicitly may be prohibitive; see Gill, Murray, Saunders, and Wright [GMSW85]. To obtain p from (1.3), the KKT-matrix is required. If H and A are sparse, they yield a sparse K. Consequently, if equations involving K are solved, it is possible to take advantage of the sparsity of the problem.

The goal of the methods described in this paper is to compute search directions directly from equations involving the KKT-matrix K or a modified KKT-matrix. We prefer to use the identical method to compute the Newton direction from (1.3) that we would use if it was known in advance that Z^THZ was positive definite. Also, a method for which there exists an efficient implementation for a large sparse matrix K is preferred. Accordingly, we consider the LBL^T-factorization, described in §3. Such a factorization computes $\Pi^TK\Pi = LBL^T$, where Π is a permutation matrix, L is a unit lower-triangular matrix, and B is a symmetric block-diagonal matrix whose diagonal blocks are of size 1×1 or 2×2. The permutations are performed in order to obtain a matrix L that is sparse and well conditioned. For unconstrained problems, the matrices K and H are identical. In this case, Moré and Sorensen [MS79] have shown how to compute a descent direction and a direction of negative curvature, whenever they exist, from L and B. We describe a pivoting strategy in the LBL^T-factorization of K, so that the ability to compute a feasible descent direction and a feasible direction of negative curvature is achieved by a single factorization of K, analogous to the method of Moré and Sorensen. The computed directions are shown to satisfy conditions needed to apply known convergence results for modified Newton methods, which state that limit points of a generated sequence of iterates satisfy the second-order necessary optimality conditions.

The above-mentioned pivoting strategy, sufficient for computing both a descent direction and a direction of negative curvature, is then shown to be necessary in order to guarantee that one is able to obtain the directions from a single factorization of K. Two other methods for computing search directions are described that do not require this pivoting strategy. One method always generates a descent direction, whereas the other method generates a search direction that is either a descent direction and/or a direction of negative curvature. Applying known convergence results for modified Newton methods, limit points of a generated sequence of iterates using these methods satisfy the first-order necessary optimality conditions. The main purpose of introducing two additional methods is that the computation of the search direction is likely to be cheaper than in the method motivated by convergence to a second-order point. The intent is that the three methods may be mixed to obtain a more efficient method.

2. Basics.

2.1. Terminology. Throughout the paper, the Hessian $\nabla^2 f(x)$ and gradient $\nabla f(x)$ are evaluated at points $x \in S(x_0)$. Given an arbitrary point in $S(x_0)$, H denotes the Hessian and g the gradient at such a point. The set of KKT-matrices for which H is evaluated at some point in $S(x_0)$ is denoted by $\mathcal{K}$. All KKT-matrices considered belong to $\mathcal{K}$. For an iterative sequence $\{x_k\}_{k=0}^{\infty}$ a subscript k is included, so that $H_k = \nabla^2 f(x_k)$ and $g_k = \nabla f(x_k)$. Also, for vectors and matrices, the norm used is always the Euclidean vector norm and the corresponding induced matrix norm. For a symmetric matrix M, we denote by $\lambda_{\min}(M)$ the smallest eigenvalue of M. The vector e_i denotes the ith unit vector of the appropriate dimension. In a number of lemmas, matrices of zero dimension may arise. In such circumstances, there is no loss of generality if we make the assumption that a matrix of zero dimension is positive definite and has norm zero.

2.2. The inertia of a matrix. Let M be any symmetric matrix. We denote by $i^+(M)$, $i^-(M)$, and $i^0(M)$, respectively, the (nonnegative) numbers of positive, negative, and zero eigenvalues of M. The *inertia* of M—denoted by $\mathrm{In}(M)$—is the associated integer triple (i^+, i^-, i^0). The following lemma states an important relationship between the inertia of the KKT-matrix K and the reduced Hessian Z^THZ.

LEMMA 2.1. *If* $\mathrm{rank}(A) = r$, *and* Z *is a matrix whose columns form a basis for the null space of* A, *then* $\mathrm{In}(K) = \mathrm{In}(Z^THZ) + (r, r, m - r)$.

Proof. See Gould [GOU85, Lemma 3.4] for the proof. □

This lemma implies that K is singular only if Z^THZ is singular or A is rank deficient. In this paper, we assume that A has full row rank, so that singularity of K always means singularity of Z^THZ. The following lemma shows that if A has full row rank, there is a uniform relationship between the nonsingularity of K and Z^THZ.

LEMMA 2.2. *Assume that* $\mathrm{rank}(A) = m$, *and let* Z *denote a matrix whose columns form a basis for the null space of* A. *For a given positive constant* c_1, *there exists a positive constant* c_2 *such that for any KKT-matrix* $K \in \mathcal{K}$ *having its smallest singular value greater than* c_1, *the smallest singular value of* Z^THZ *is greater than* c_2.

Proof. If A has full row rank, Lemma 2.1 implies that K is nonsingular if and only if Z^THZ is nonsingular. For a symmetric matrix, the singular values are the absolute values of the eigenvalues. The norm of K is bounded, since $S(x_0)$ is compact. The result follows by observing that the eigenvalues of K and Z^THZ vary continuously with H. □

3. The LBL^T-factorization.

3.1. The factorization.

An efficient method for solving equations involving symmetric indefinite matrices is to use the factorization

$$\Pi^TK\Pi = LBL^T, \tag{3.1}$$

where Π is a permutation matrix that represents column interchanges in K, L is a unit lower-triangular matrix, and B is a symmetric block-diagonal matrix whose blocks are of size 1×1 or 2×2. We shall refer to this factorization as the LBL^T-factorization. Various algorithms for computing the factors have been proposed; see Bunch and Parlett [BP71] and Bunch and Kaufman [BK77].

The algorithms proposed by Bunch and Parlett [BP71] and Bunch and Kaufman [BK77] have the property that the 2×2 blocks are always nonsingular with one positive and one negative eigenvalue. Other algorithms have been proposed where the 2×2 blocks may have both eigenvalues of the same sign; see, for example, Duff and Reid [DR82], [DR83].

In this paper, we do not elaborate on which algorithm to use, except to assume that the algorithm performs a *regular* LBL^T-factorization, where we define an LBL^T-factorization to be regular if it satisfies (i) the norm of L is bounded, and (ii) the 2×2 blocks always have one positive and one negative eigenvalue. Property (i) is essential in our analysis, whereas property (ii) is used only to simplify the notation. The algorithm of Bunch and Parlett [BP71] is an example of an algorithm that yields a regular LBL^T-factorization.

Forming the LBL^T-factorization may be viewed as a step-wise procedure that repeatedly computes *Schur complements* of decreasing dimension. If K is partitioned as

$$K = \begin{pmatrix} T & N^T \\ N & G \end{pmatrix}, \tag{3.2}$$

and T is nonsingular, the *Schur complement of* T *in* K will be denoted by K/T and is defined as

$$K/T = G - NT^{-1}N^T.$$

The matrix K/T will be referred to as "the" Schur complement when the matrix T is clear from the context. By convention, we define the Schur complement as K when T has dimension zero.

Given the partition from (3.2), the identity

$$\begin{pmatrix} T & N^T \\ N & G \end{pmatrix} = \begin{pmatrix} I & 0 \\ NT^{-1} & I \end{pmatrix} \begin{pmatrix} T & 0 \\ 0 & K/T \end{pmatrix} \begin{pmatrix} I & T^{-1}N^T \\ 0 & I \end{pmatrix} \tag{3.3}$$

holds. In the first step of the factorization, the matrix T is a 1×1 or 2×2 principal submatrix of K. Symmetric row and column interchanges may be necessary in order to obtain a T that is suitable as *pivot*. The elimination step yields one or two rows of L^T as $(I \quad T^{-1}N^T)$ and a diagonal block of B of size 1×1 or 2×2 from T. In the next step, the process is repeated for the Schur complement K/T. Eventually, the Schur complement has dimension zero, and the algorithm terminates with permutation matrix Π, unit lower-triangular matrix L, and block-diagonal matrix B. Algorithms differ in the way the pivot is chosen. Since L is a unit lower-triangular matrix, when the norm of L is bounded, then the norm of L^{-1} is also bounded. Consequently, it follows from (3.1) that K is arbitrarily close to a singular matrix if and only if B is arbitrarily close to a singular matrix. In our applications, the inertia of K is required. This inertia is readily available if B is known, since Sylvester's law of inertia yields $\mathrm{In}(K) = \mathrm{In}(B)$; see [GV89, p. 416].

We require bounds on the relative magnitude of the smallest eigenvalues of B and LBL^T. In the following lemma, suitable bounds are derived from the proof of Sylvester's law of inertia that is given in Golub and Van Loan [GV89, p. 416].

LEMMA 3.1. *Let* L *be an* $n\times n$ *nonsingular matrix and let* B *be an* $n\times n$ *symmetric matrix with at least one nonpositive eigenvalue. It follows that*

$$\|L^{-1}\|^2 \lambda_{\min}(LBL^T) \le \lambda_{\min}(B) \le \frac{1}{\|L\|^2} \lambda_{\min}(LBL^T).$$

Proof. By definition

$$\lambda_{\min}(B) = \min_{x\neq 0} \frac{x^T Bx}{x^T x}.$$

Let y denote an eigenvector corresponding to the smallest eigenvalue of B. Upon observing that L is nonsingular we obtain

$$\lambda_{\min}(B) = \frac{y^TBy}{y^Ty} = \frac{y^TL^{-1}LBL^TL^{-T}y}{y^TL^{-1}L^{-T}y} \cdot \frac{y^TL^{-1}L^{-T}y}{y^Ty}. \tag{3.4}$$

The right-hand side of (3.4) is a product of two factors. Since the second factor is positive, and $\lambda_{\min}(B)$ is nonpositive, it follows that the first factor is nonpositive. Consequently, a lower bound on the smallest eigenvalue of B is given by

$$\lambda_{\min}(B) \ge \min_{x\neq 0} \frac{x^TL^{-1}LBL^TL^{-T}x}{x^TL^{-1}L^{-T}x} \cdot \max_{x\neq 0} \frac{x^TL^{-1}L^{-T}x}{x^Tx}. \tag{3.5}$$

It follows from (3.5) that

$$\lambda_{\min}(B) \geq \lambda_{\min}(LBL^T) \cdot \|L^{-1}\|^2. \tag{3.6}$$

Since $0 \geq \lambda_{\min}(B)$, it follows from (3.6) that LBL^T has at least one nonpositive eigenvalue. Therefore, B may be replaced by LBL^T in (3.6). If we also replace L in (3.6) by L^{-1} we get

$$\lambda_{\min}(LBL^T) \geq \lambda_{\min}(L^{-1}LBL^TL^{-T}) \cdot \|L\|^2 = \lambda_{\min}(B) \cdot \|L\|^2. \tag{3.7}$$

The required result now follows by combining (3.6) and (3.7). □

3.2. Computational considerations. When K is sparse, the LBL^T-factorization of K may be carried out in two steps. First, in the *analyze phase*, a symbolic factorization based on the nonzero elements in K is performed. This symbolic factorization yields an ordering of the rows and columns that attempts to minimize the number of nonzeros in L. Then, in the *numerical phase*, the factors are computed, attempting to maintain the ordering given by the analyze phase. However, for numerical reasons, it may be necessary to perform additional permutations in the numerical phase. Consequently, the permutation matrix Π in (3.1) is the combination of the ordering from the analyze phase and the additional permutations from the numerical phase. In a modified Newton method, where a sequence of K-matrices is factorized, the analyze phase need only be performed once, since the positions of the nonzero elements in the Hessian remain the same. A robust and efficient routine that performs such a two-step factorization is the Harwell routine MA27; see Duff and Reid [DR82], [DR83]. (MA27 may use 2 × 2 pivots that are not indefinite. To fit the discussion here, such a pivot may be viewed equivalent to two consecutive 1 × 1 pivots.)

3.3. Definition of pivot types. In a regular LBL^T-factorization, the matrix B consists of diagonal blocks of size 1 × 1 or 2 × 2, where the 2 × 2-blocks are indefinite. These diagonal blocks are the pivots defined by the factorization algorithm. It is of importance to distinguish between different types of blocks.

A 1 × 1-block is defined to be of type H^+ if it is positive, and if the element in the same position in $\Pi^T K \Pi$ is a diagonal element of H. Similarly, 1 × 1-blocks are defined to be of type H^- and H^0 if they satisfy the same position requirements as the blocks of type H^+, but are negative or zero, respectively. We denote by n_H^+, n_H^-, and n_H^0, respectively, the number of such 1 × 1-blocks.

A 1 × 1-block is defined to be of type A^+ if it is positive, and if the element in the same position in $\Pi^T K \Pi$ is a diagonal element of the zero-part of K. Similarly, 1 × 1-blocks are defined to be of type A^- and A^0 if they satisfy the same position requirements as the blocks of type A^+, but are negative or zero, respectively. We denote by n_A^+, n_A^-, and n_A^0, respectively, the number of such 1 × 1-blocks.

A 2 × 2-block is defined to be of type *HH* if the elements in $\Pi^T K \Pi$ with the same position are elements of H. We denote by n_{HH} the number of such 2 × 2-blocks.

A 2 × 2-block is defined to be of type *AA* if the elements in $\Pi^T K \Pi$ with the same position are elements of the zero-part of K. We denote by n_{AA} the number of such 2 × 2-blocks.

A 2 × 2-block is defined to be of type *HA* if the elements in $\Pi^T K \Pi$ with the same position consist of one diagonal element from H, one diagonal element from the zero-part of K, and two elements from A.

3.4. Applications in unconstrained minimization. For unconstrained problems, Moré and Sorensen [MS79] have shown how to compute a descent direction s and a direction of negative curvature d, whenever such directions exist, using the LBL^T-factorization of the Hessian H, given by $\Pi^T H \Pi = LBL^T$. The descent direction s satisfies the equation $\Pi L\bar{B}L^T\Pi^T s = -g$, where $\bar{B}$ is a positive definite modification of B. If H is sufficiently positive definite, then $\bar{B} = B$ and s is the Newton search direction. If H is not positive definite, $\bar{B}$ is obtained as $\bar{B} = B + D$, where D is a block-diagonal matrix with the same block structure as B, and D has rank equal to the number of nonpositive eigenvalues of H. If H has at least one negative eigenvalue, the direction of negative curvature, d, satisfies the equation $L^T\Pi^T d = \pm\sqrt{-\lambda_{\min}(B)}v$, where v is an eigenvector of unit norm corresponding to $\lambda_{\min}(B)$, the smallest eigenvalue of B. The sign is chosen so that $g^Td \le 0$. If H has all eigenvalues nonnegative, $d = 0$. For details, see Moré and Sorensen [MS79].

Similarly, in the linear equality-constrained case, a feasible descent direction (assuming such a direction exists), may be obtained by solving

$$\begin{pmatrix} \bar{H} & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} s \\ -\mu \end{pmatrix} = \begin{pmatrix} -g \\ 0 \end{pmatrix}, \tag{3.8}$$

where $\bar{H}$ is a modification of H such that $Z^T\bar{H}Z$ is positive definite. Utilizing the idea of Moré and Sorensen [MS79] from the unconstrained case, we may try to use the LBL^T-factorization of K to modify blocks of B of types H^-, H^0, and HH, yielding a matrix $\bar{B}$ such that $L\bar{B}L^T$ has n positive eigenvalues and m negative eigenvalues. The following lemma shows that there is a sufficient number of blocks of types H^-, H^0, and HH to create such a matrix $\bar{B}$.

LEMMA 3.2. *If* $\operatorname{In}(K) = (i^+, i^-, i^0)$, *then* $n_H^- + n_H^0 + n_{HH} \ge i^- + i^0 - m$.

Proof. Assume the contrary, that $n_H^- + n_H^0 + n_{HH} < i^- + i^0 - m$. The dimensions of H and A imply

$$n_H^+ + n_H^- + n_H^0 + 2n_{HH} + n_{HA} = n, \tag{3.9a}$$

$$n_A^+ + n_A^- + n_A^0 + n_{HA} + 2n_{AA} = m. \tag{3.9b}$$

If $\operatorname{In}(K) = (i^+, i^-, i^0)$, we get

$$n_H^+ + n_{HH} + n_{HA} + n_A^+ + n_{AA} = i^+, \tag{3.10a}$$

$$n_H^- + n_{HH} + n_{HA} + n_A^- + n_{AA} = i^-, \tag{3.10b}$$

$$n_H^0 + n_A^0 = i^0. \tag{3.10c}$$

Adding (3.10b) and (3.10c), we obtain $n_{HA} + n_A^- + n_{AA} + n_A^0 > m$, contradicting (3.9b). Thus $n_H^- + n_H^0 + n_{HH} < i^- + i^0 - m$ cannot hold. □

However, although there are sufficiently many block elements of types H^-, H^0, and HH, if no additional conditions are imposed on Π, the ordering may be insufficient to guarantee that only the H-part of K is altered when a block element of type H^-, H^0, or HH is modified. Consequently, if an equation involving $L\bar{B}L^T$ is solved, there is no straightforward analogous way of obtaining a descent direction s such that $As = 0$. As an example, consider

$$H = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 2 & 1 \end{pmatrix}, \quad \text{so that} \quad K = \begin{pmatrix} 2 & 0 & 2 \\ 0 & -2 & 1 \\ 2 & 1 & 0 \end{pmatrix}.$$

If Π is the identity matrix, we obtain

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -\frac{1}{2} & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -\frac{3}{2} \end{pmatrix}.$$

The matrix B has $n_H^+ = n_H^- = n_A^- = 1$. Analogous to the unconstrained case, we make the second diagonal element of B positive by adding a positive number d_{22}. However, the matrix $L\bar{B}L^T$ will differ from LBL^T in the A-part and the zero-part, since

$$L(B+D)L^T = \begin{pmatrix} 2 & 0 & 2 \\ 0 & -2 & 1 \\ 2 & 1 & 0 \end{pmatrix} + d_{22} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & \frac{1}{4} \end{pmatrix}.$$

Because of the lower-triangular structure of L, if a diagonal block of B is altered, the modification may alter elements of the permuted K with greater row and column numbers. Consequently, if there are rows of A with greater row numbers, there is a danger of altering parts of K other than the H-part.

It may appear that, by analogy with the unconstrained case, a feasible direction of negative curvature may be obtained by solving the equation $L^T\Pi^T u = v$, where v is an eigenvector of B corresponding to a negative eigenvalue of a block of type H^- or HH. (Note that such a vector v always exists if the reduced Hessian has at least one negative eigenvalue, since Lemma 2.1 implies that K has more than m negative eigenvalues. Consequently, since there are at most m block elements of types A^-, HA, and AA, there must exist a block element of type H^- or HH for K to have more that m negative eigenvalues.) The vector d is then defined as the first n components of u. However, the ordering of the rows and columns of K given by Π may be insufficient to guarantee that $Ad = 0$. In our example matrix, we obtain $d = (0\ 1)^T$ for $v = (0\ 1\ 0)^T$, and consequently, $Ad = 1 \neq 0$.

Conn and Gould [CG84] have proposed an algorithm for finding feasible directions of negative curvature based on the LBL^T-factorization of K. In addition to the original factorization, the algorithm requires the factorization of a triangle-like matrix of dimension $m \times m$. In this paper, we adopt a different view and investigate what pivoting strategy is necessary and sufficient to obtain the desired search directions in a single factorization of K, using L and B in an analogous way to the approach of Moré and Sorensen [MS79] for the unconstrained case.

4. Sufficient pivot conditions. If K is factorized using a regular LBL^T-factorization, the permutations are performed in order to obtain sparsity of the factors and boundedness of $\|L\|$. In the example given in §3.4, it was shown that these permutations may be insufficient for computing a feasible descent direction and a feasible direction of negative curvature. In this section, we investigate what conditions may be imposed on the pivots in order to ensure the ability to compute these directions, using L and B, whenever the directions exist.

Upon completion of the factorization of K, we have computed a permutation matrix Π, a lower-triangular matrix L and a block-diagonal matrix B such that $LBL^T = \Pi^T K \Pi$. We will consider a specific step in the factorization. Therefore, the permutation matrix Π is partitioned as

$$\Pi = \begin{pmatrix} \Pi_1 & \Pi_2 \end{pmatrix}, \tag{4.1}$$

where the matrix $\Pi_1^T K \Pi_1$ is the principal submatrix of K for which the factors have been computed. Consequently, $\Pi_1^T K \Pi_1 = L_{11} B_1 L_{11}^T$, where L_{11} and B_1 are the leading principal submatrices of L and B, respectively. Note that at this step of the factorization, the matrix Π_2 has not yet been determined.

The matrix $\Pi_1^T K \Pi_1$ is a principal submatrix of K, and we introduce a permutation $\bar{\Pi}$ for which

$$\bar{\Pi}^T K \bar{\Pi} = \begin{pmatrix} H_{11} & H_{12} & A_{11}^T & A_{21}^T \\ H_{21} & H_{22} & A_{12}^T & A_{22}^T \\ A_{11} & A_{12} & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 \end{pmatrix}.$$

The matrix $\bar{\Pi}$ and the partition of $\bar{\Pi}^T K \bar{\Pi}$ are such that if we define K_{11} as

$$K_{11} = \begin{pmatrix} H_{11} & A_{11}^T \\ A_{11} & 0 \end{pmatrix}, \tag{4.2}$$

the matrix $\Pi_1^T K \Pi_1$ is a permuted version of K_{11}. The matrix K_{11} is a principal submatrix of K such that H_{11} is an $n_1 \times n_1$ principal submatrix of H, and A_{11} is an $n_1 \times m_1$ submatrix of A. To simplify the notation, we say that K_{11} *contains* n_1 rows of H and m_1 rows of A.

Since $\Pi_1^T K \Pi_1$ denotes the part of K that has been factorized at a particular step, its size increases from step to step. We say that $\Pi_1^T K \Pi_1$ and the associated K_{11} are *expanded* when one step in the factorization is performed. For example, if in one step of the factorization, a pivot of type HA is selected, the size of $\Pi_1^T K \Pi_1$ and K_{11} is increased by two, and both n_1 and m_1 are increased by one.

It is shown in this section that by selecting sufficiently nonsingular pivots of types H^+, A^-, and HA, until the corresponding K_{11} contains all rows of A, the ability to compute a feasible descent direction and a feasible direction of negative curvature is achieved. A matrix K_{11} created in such a way will prove crucial for the computation of descent directions and directions of negative curvature.

Properties of $\Pi_1^T K \Pi_1$ that we want to describe are also properties of K_{11}. In particular, the eigenvalues of K_{11} and $\Pi_1^T K \Pi_1$ are identical. We consider K_{11} since its properties are easier to describe. The following lemma gives an explicit representation of K_{11}^{-1}.

LEMMA 4.1. *If K_{11} is nonsingular, then*

$$K_{11}^{-1} = \begin{pmatrix} H_{11} & A_{11}^T \\ A_{11} & 0 \end{pmatrix}^{-1} = \begin{pmatrix} \bar{Z}(\bar{Z}^T H_{11} \bar{Z})^{-1} \bar{Z}^T & \bar{Y} \\ \bar{Y}^T & -\bar{Y}^T H_{11} \bar{Y} \end{pmatrix},$$

where $\bar{Z}$ is an orthonormal matrix whose columns form a basis for the null space of A_{11}, and the matrix $\bar{Y}$ satisfies the matrix equation

$$\begin{pmatrix} H_{11} & A_{11}^T \\ A_{11} & 0 \end{pmatrix} \begin{pmatrix} \bar{Y} \\ -\Lambda \end{pmatrix} = \begin{pmatrix} 0 \\ I \end{pmatrix}, \tag{4.3}$$

where $\Lambda = \bar{Y}^T H_{11} \bar{Y}$. If A_{11} is square, the matrix $\bar{Z}(\bar{Z}^T H_{11} \bar{Z})^{-1} \bar{Z}^T$ is to be interpreted as a zero matrix of dimension $m_1 \times m_1$.

Proof. Direct substitution in (4.3) shows that $\Lambda = \bar{Y}^T H_{11} \bar{Y}$. Lemma 2.1 shows that nonsingularity of K_{11} yields nonsingularity of $\bar{Z}^T H_{11} \bar{Z}$. Thus it remains to verify that

$$\begin{pmatrix} H_{11} & A_{11}^T \\ A_{11} & 0 \end{pmatrix} \begin{pmatrix} \bar{Z}(\bar{Z}^T H_{11} \bar{Z})^{-1} \bar{Z}^T \\ \bar{Y}^T \end{pmatrix} = \begin{pmatrix} I \\ 0 \end{pmatrix}. \tag{4.4}$$

This holds if and only if $H_{11}\bar{Z}(\bar{Z}^T H_{11}\bar{Z})^{-1}\bar{Z}^T + A_{11}^T \bar{Y}^T = I$. If A_{11} is square, then $\bar{Y} = A_{11}^{-1}$, and (4.4) holds for $\bar{Z}(\bar{Z}^T H_{11}\bar{Z})^{-1}\bar{Z}^T = 0$. Otherwise, (4.3) gives $\bar{Z}^T H_{11} \bar{Y} = 0$, and it follows that

$$\begin{pmatrix} \bar{Z}^T \\ \bar{Y}^T \end{pmatrix} \begin{pmatrix} H_{11}\bar{Z}(\bar{Z}^T H_{11}\bar{Z})^{-1}\bar{Z}^T + A_{11}^T \bar{Y}^T \end{pmatrix} = \begin{pmatrix} \bar{Z}^T \\ \bar{Y}^T \end{pmatrix}. \tag{4.5}$$

The proof is now completed by showing that $(\bar{Z} \quad \bar{Y})$ is nonsingular. Assume that $\bar{Z} v_z + \bar{Y} v_y = 0$. Premultiplication by A_{11} yields $v_y = 0$. Since the columns of $\bar{Z}$ are linearly independent, $v_z = 0$. □

The following lemma shows that when $\Pi_1^T K \Pi_1$ has inertia $(n_1, m_1, 0)$, a row in $\Pi^T K \Pi / \Pi_1^T K \Pi_1$ corresponding to a row of A cannot be almost linearly dependent on the other rows of $\Pi^T K \Pi / \Pi_1^T K \Pi_1$ unless A is almost rank deficient.

LEMMA 4.2. *For given positive constants c_1 and M, let K be any KKT-matrix in $\mathcal{K}$ with a principal submatrix K_{11} such that* $\mathrm{In}(K_{11}) = (n_1, m_1, 0)$, *where the smallest singular value of K_{11} is greater than c_1. Define a matrix S, which is a permuted version of the Schur complement $\Pi^T K \Pi / \Pi_1^T K \Pi_1$, in partitioned form as*

$$\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} H_{22} & A_{22}^T \\ A_{22} & 0 \end{pmatrix}$$
$$- \begin{pmatrix} H_{21} & A_{12}^T \\ A_{21} & 0 \end{pmatrix} \begin{pmatrix} H_{11} & A_{11}^T \\ A_{11} & 0 \end{pmatrix}^{-1} \begin{pmatrix} H_{12} & A_{21}^T \\ A_{12} & 0 \end{pmatrix}.$$

If $m_1 < m$, there exists a positive constant c_2, such that if

$$- e_i^T S_{22} e_i \le \frac{\epsilon}{2Mn} \tag{4.6a}$$

and

$$\|e_i^T S_{21}\| \le \sqrt{\epsilon} \tag{4.6b}$$

for some i, then

$$\left\| \begin{pmatrix} -e_i^T A_{21} \bar{Y} & e_i^T \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \right\| \le c_2 \sqrt{\epsilon}, \tag{4.7}$$

where $\bar{Y}$ is defined in (4.3) *and ϵ is any nonnegative scalar.*

Proof. Assume that $m_1 < m$ and that there is an i such that (4.6a) and (4.6b) hold for some nonnegative ϵ. Utilizing the notation of Lemma 4.1, we obtain

$$S_{11} = H_{22} - H_{21}\bar{Z}(\bar{Z}^T H_{11}\bar{Z})^{-1}\bar{Z}^T H_{12} - A_{12}^T \bar{Y}^T H_{12} - H_{21}\bar{Y} A_{12} + A_{12}^T \bar{Y}^T H_{11} \bar{Y} A_{12},$$
$$S_{12} = A_{22}^T - H_{21}\bar{Z}(\bar{Z}^T H_{11}\bar{Z})^{-1}\bar{Z}^T A_{21}^T - A_{12}^T \bar{Y}^T A_{21}^T,$$
$$S_{21} = A_{22} - A_{21}\bar{Z}(\bar{Z}^T H_{11}\bar{Z})^{-1}\bar{Z}^T H_{12} - A_{21}\bar{Y} A_{12},$$
$$S_{22} = -A_{21}\bar{Z}(\bar{Z}^T H_{11}\bar{Z})^{-1}\bar{Z}^T A_{21}^T.$$

The definition of S_{22} yields

$$-e_i^T S_{22} e_i = e_i^T A_{21} \bar{Z}(\bar{Z}^T H_{11} \bar{Z})^{-1} \bar{Z}^T A_{21}^T e_i.$$

If the smallest singular value of K_{11} is greater than a positive constant c_1 and $\operatorname{In}(K_{11}) = (n_1, m_1, 0)$, Lemmas 2.1 and 2.2 guarantee that $\bar{Z}^T H_{11} \bar{Z}$ is positive definite with its smallest eigenvalue greater than a positive constant. By the assumed compactness of $S(x_0)$, the elements in H_{11} are bounded in magnitude. Consequently, there exists a positive constant $\bar{c}_1$, such that

$$\frac{\epsilon}{2Mn} \geq -e_i^T S_{22} e_i \geq \bar{c}_1 \|\bar{Z}^T A_{21}^T e_i\|^2, \tag{4.8}$$

where the left-hand inequality follows from the assumption that (4.6a) holds. Using the definition of S_{21} we obtain

$$e_i^T A_{22} - e_i^T A_{21} \bar{Y} A_{12} = e_i^T S_{21} + e_i^T A_{21} \bar{Z}(\bar{Z}^T H_{11} \bar{Z})^{-1} \bar{Z}^T H_{12}.$$

It follows from this equation, (4.6b), (4.8), the nonsingularity of K_{11}, and the boundedness of $\|H\|$ that there exists a positive constant $\bar{c}_2$, such that

$$\|e_i^T A_{22} - e_i^T A_{21} \bar{Y} A_{12}\| \leq \bar{c}_2 \sqrt{\epsilon}. \tag{4.9}$$

The identities $A_{11}\bar{Y} = I$ and $A_{21}^T e_i = A_{11}^T (A_{11} A_{11}^T)^{-1} A_{11} A_{21}^T e_i + \bar{Z}\bar{Z}^T A_{21}^T e_i$ yield

$$e_i^T A_{21} \bar{Y} A_{11} = e_i^T A_{21} - e_i^T A_{21} \bar{Z}\bar{Z}^T + e_i^T A_{21} \bar{Z}\bar{Z}^T \bar{Y} A_{11}.$$

Utilizing (4.8) and the nonsingularity of K_{11} we obtain

$$\|e_i^T A_{21} - e_i^T A_{21} \bar{Y} A_{11}\| \leq \bar{c}_3 \sqrt{\epsilon} \tag{4.10}$$

for some positive constant $\bar{c}_3$. Consequently, (4.9) and (4.10) yield the existence of a positive constant c_2, such that

$$\left\| \begin{pmatrix} -e_i^T A_{21} \bar{Y} & e_i^T \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \right\| \leq c_2 \sqrt{\epsilon}. \qquad \square$$

In the following lemma, we use the full-rank assumption on A and Lemma 4.2 to show that if K_{11} has exactly m_1 ($m_1 < m$) negative eigenvalues and has its smallest singular value greater than a positive constant, it is always possible to expand K_{11} so that m_1 is increased by one.

LEMMA 4.3. *For a given positive constant c_1, let K be any KKT-matrix in $\mathcal{K}$ with a principal submatrix K_{11}, such that* $\operatorname{In}(K_{11}) = (n_1, m_1, 0)$, *where the smallest singular value of K_{11} is greater than c_1. Assume that the factorization of $\Pi_1^T K \Pi_1$ is known, and that the next pivot is to be chosen from the Schur complement $\Pi^T K \Pi / \Pi_1^T K \Pi_1$. If $m_1 < m$, there exists a positive constant c_3, such that there is a pivot in $\Pi^T K \Pi / \Pi_1^T K \Pi_1$ of type A^- or HA whose determinant is less than $-c_3$.*

Proof. Let c_1 be a given positive constant, and K any KKT-matrix in $\mathcal{K}$ with a principal submatrix K_{11} such that $\operatorname{In}(K_{11}) = (n_1, m_1, 0)$, where the smallest singular value of K_{11} is greater than c_1. Using the terminology of Lemma 4.2, if K_{11} has inertia $(n_1, m_1, 0)$ and the smallest singular value greater than c_1, Lemma 4.1 and the definition of S guarantee the existence of a constant M such that $\|S\| \leq M$. Let $e_i^T(S_{21} \quad S_{22})$ be a row of S corresponding to a row of A and let l denote the corresponding row number of A. Since A has full row rank, there exists a positive constant $\bar{c}$ such that

$$\left\| e_l^T A - \sum_{j \neq l} \alpha_j e_j^T A \right\| \geq \bar{c} \tag{4.11}$$

for all scalars α_j. Given c_1 and M, let c_2 denote the constant from Lemma 4.2. Let $\bar{\epsilon}$ denote the positive constant defined by $c_2\sqrt{\bar{\epsilon}} = \bar{c}/2$. For this choice of $\bar{\epsilon}$, at least one of

$$-e_i^T S_{22} e_i > \frac{\bar{\epsilon}}{2Mn} \tag{4.12a}$$

and

$$\|e_i^T S_{21}\| > \sqrt{\bar{\epsilon}} \tag{4.12b}$$

must hold since, if (4.12a) and (4.12b) do not hold, Lemma 4.2 implies that there exists a vector α such that

$$\left\| e_l^T A - \sum_{j \neq l} \alpha_j e_j^T A \right\| \leq c_2\sqrt{\bar{\epsilon}}, \tag{4.13}$$

which contradicts (4.11). It is now shown that if (4.12a) holds, there exists a pivot of type A^-, and if (4.12a) does not hold but (4.12b) holds, there exists a suitable pivot of type HA. Assume that (4.12a) holds. Then $e_i^T S_{22} e_i$ is a pivot of type A^- whose determinant is less than $-\bar{\epsilon}/(2Mn)$. Assume that (4.12b) holds but (4.12a) does not hold. If (4.12b) holds, it must hold that $n - n_1 > 0$, since the length of the row vector $e_i^T S_{21}$ is $n - n_1$. Moreover, there must exist a j, such that

$$|e_i^T S_{21} e_j| \geq \sqrt{\frac{\bar{\epsilon}}{n - n_1}}. \tag{4.14}$$

For this j, let P denote the 2×2 matrix defined as

$$P = \begin{pmatrix} e_j^T S_{11} e_j & e_j^T S_{12} e_i \\ e_i^T S_{21} e_j & e_i^T S_{22} e_i \end{pmatrix}.$$

By assumption, (4.12a) does not hold, and consequently, since $e_i^T S_{22} e_i \leq 0$, it holds that $|e_i^T S_{22} e_i| \leq \bar{\epsilon}/(2Mn)$. This assumption and the inequality $|e_j^T S_{11} e_j| \leq M$ yield

$$\det(P) \leq \frac{\bar{\epsilon}}{2n} - |e_i^T S_{21} e_j|^2.$$

It follows from (4.14) that $|e_i^T S_{21} e_j| \geq \sqrt{\bar{\epsilon}/n}$, and consequently

$$\det(P) \leq -\frac{\bar{\epsilon}}{2n}.$$

The determinant of P is negative, and therefore P must have one positive and one negative eigenvalue. Consequently, P is a pivot of type HA whose determinant is less than a negative constant. Therefore, there exists a positive constant c_3 for which there is a pivot of type A^- or HA whose determinant is less than $-c_3$. □

Based on this lemma, the factorization algorithm is stated in Algorithm 4.1. Initially, the matrix K_{11} has dimension zero, and for such a matrix the assumptions of Lemma 4.3 hold. It follows from Lemma 4.3 that there exists a constant pivot tolerance tol so that there is a pivot P of type H^+, A^-, or HA for which $|\det(P)| \geq$ tol. K_{11} is expanded using this pivot. This process is repeated until K_{11} contains all rows of A. When K_{11} contains all rows of A it is considered the "final" K_{11}-matrix, and the remaining Schur complement is factorized using a regular LBL^T-factorization.

ALGORITHM 4.1. *An inertia-controlling LBL^T-factorization of K.*

```
tol ← pivot tolerance;
repeat
    ex ← pivot $P$ of type $H^+$, $A^-$, or $HA$ exists with $|\det(P)| \geq$ tol;
    if ex then
        pivot ← $P$;
    else
        Set tol to the largest value for which a pivot $P$ of type $H^+$, $A^-$, or
            $HA$ exists with $|\det(P)| =$ tol;
        pivot ← $P$;
    end if
    Expand $K_{11}$ using $P$ as pivot;
    Update $m_1$ and $n_1$;
until $m_1 = m$
Factorize the remainder of $K$ using a regular $LBL^T$-algorithm;
```

In order to show that this algorithm is well defined, it is essential to show that the required properties of K_{11} are maintained when K_{11} is expanded. The following lemma shows that this is true.

LEMMA 4.4. *For a given positive constant c_1, let K be any KKT-matrix in $\mathcal{K}$ with a principal submatrix K_{11} such that* $\text{In}(K_{11}) = (n_1, m_1, 0)$ *and the smallest singular value of K_{11} is greater than c_1. Assume that the factorization of $\Pi_1^T K \Pi_1$ is known, and that the next pivot P is of type H^+, A^-, or HA and has $|\det(P)| \geq c_3$, where c_3 is a positive constant. Then, there exists a positive constant c_4 such that if K_{11} is expanded using P as pivot, the expanded K_{11} has inertia $(n_1, m_1, 0)$ and its smallest singular value is greater than c_4.*

Proof. If K_{11} is expanded using a pivot P of type H^+, A^-, or HA, the expanded matrix still has inertia $(n_1, m_1, 0)$. The existence of c_4 remains to be established. Let $\bar{K}_{11}$ denote the expanded K_{11}-matrix. Since the smallest singular value of K_{11} is greater than c_1, there must exist a positive constant $\bar{c}_1$ such that $|\det(K_{11})| \geq \bar{c}_1$. It follows from (3.3) that $\det(\bar{K}_{11}) = \det(K_{11})\det(P)$. If $|\det(P)| \geq c_3$, it follows that $|\det(\bar{K}_{11})| \geq \bar{c}_1 c_3$. Consequently, since the norm of K is bounded, there exists a positive constant c_4 such that the smallest singular value of $\bar{K}_{11}$ is greater than c_4. □

This lemma shows that, since A has full row rank, a K_{11}-matrix with inertia $(n_1, m_1, 0)$ and the smallest singular value greater than a positive constant may be expanded until it contains all rows of A and has m negative eigenvalues. Moreover, the smallest singular value of the final K_{11}-matrix is bounded away from zero. Note also that the pivot tolerance tol is bounded away from zero by a constant, since it is greater than a positive constant for each step in the expansion of K_{11}.

If it is required that only sufficiently nonsingular pivots of types H^+, A^-, and HA are utilized, until a matrix $\Pi_1^T K \Pi_1$ containing all rows of A is created, this will impose other permutations than the ones provided in a regular algorithm for the LBL^T-factorization. Consequently, it is essential that the boundedness of the norm of the L-matrix is maintained. Since the pivots of the associated B_1-matrix all have singular values bounded away from zero by a constant, the following lemma shows that the boundedness of the norm of L is maintained.

LEMMA 4.5. *Assume that* $\Pi_1^T K \Pi_1 = L_{11} B_1 L_{11}^T$, *with its smallest singular value greater than a positive constant. Furthermore, assume that* $\|L_{11}\|$ *is bounded by a constant. For any* Π_2 *consistent with* Π_1, *defined from* (4.1), *let* L_{21} *denote the unique solution to the matrix equation* $L_{21} B_1 L_{11}^T = \Pi_2^T K \Pi_1$, *and assume that* $\|L_{21}\|$ *is bounded by a constant. If* K_{11} *is expanded by a pivot that has the absolute value of its determinant bounded away from zero by a positive constant, then the expanded* L_{11} *and* L_{21} *still have norm bounded by a constant.*

Proof. Let P denote the pivot, which is of size 1×1 or 2×2. Since K and $(\Pi_1^T K \Pi_1)^{-1}$ have bounded norm, it follows that the norm of $\Pi^T K \Pi / \Pi_1^T K \Pi_1$ is bounded. Consequently, the norm of P is bounded by a constant. If, in addition, the determinant of P is bounded away from zero by a constant in absolute value, the norm of P^{-1} is bounded by a constant. Comparing with (3.3), it follows that L_{11} is expanded with a zero block at the upper-right corner, a 1×1 or 2×2 unit matrix at the lower-right corner, and one or two rows from the previous L_{21} at the lower-left corner. Consequently, the expanded L_{11} has bounded norm. Similarly, it follows from (3.3) that one or two rows are removed from L_{21}, and that one or two columns are added to L_{21} by postmultiplying one or two columns of $\Pi^T K \Pi / \Pi_1^T K \Pi_1$ by P^{-1}. Since P^{-1} has bounded norm, it follows that the norm of the expanded L_{21} is bounded. □

When a matrix K_{11} that contains all rows of A has been created by accepting pivots of types H^+, A^-, and HA, as described in Algorithm 4.1, it is considered the final K_{11}-matrix. For the rest of this section, the matrix K_{11} is assumed to be such a final K_{11}-matrix. The factors from the factorization described in Algorithm 4.1 are also assumed to be known. For this final K_{11}-matrix, the partition of Π from (4.1) is used. This partition induces a partition of L, B, and $\Pi^T K \Pi$ as

$$(4.15)\quad \begin{pmatrix} \Pi_1^T K \Pi_1 & \Pi_1^T K \Pi_2 \\ \Pi_2^T K \Pi_1 & \Pi_2^T K \Pi_2 \end{pmatrix} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix} \begin{pmatrix} L_{11}^T & L_{21}^T \\ 0 & L_{22}^T \end{pmatrix}.$$

If K_{11} contains all rows of A, all blocks in B_2 correspond to pivots of type H^+, H^-, or HH. This means that modifications of B_2 only alter the H-part of K. If the factorization of K given in Algorithm 4.1 is known, the following theorem shows how to obtain a feasible descent direction.

THEOREM 4.6. *Let* K *be any KKT-matrix in* $\mathcal{K}$. *Assume that the* LBL^T*- factorization of* K *given in Algorithm* 4.1 *is known. Assume that the eigenvalue decomposition of* B_2 *is known as* $B_2 = U \Lambda U^T$. *For a positive tolerance* c_{TOL}, *let the ith diagonal element of the matrix* $\bar{\Lambda}$, *denoted by* $\bar{\lambda}_i$, *be defined using the ith diagonal element of* Λ, *denoted by* λ_i, *as* $\bar{\lambda}_i = \max\{|\lambda_i|, c_{\text{TOL}}\}$, *and let* $\bar{B}_2 = U \bar{\Lambda} U^T$. *Let* $\bar{B}_1 = B_1$ *and let* $\bar{B}$ *denote the matrix consisting of the diagonal blocks* $\bar{B}_1$ *and* $\bar{B}_2$. *Let* $z = (s^T \ \ \mu^T)^T$, *where* s *is an n-vector and* μ *is an m-vector, be defined as* $z = \Pi \tilde{z}$, *where* $\tilde{z}$ *satisfies the equation*

$$L \bar{B} L^T \tilde{z} = \Pi^T \begin{pmatrix} -g \\ 0 \end{pmatrix}.$$

Then, $As = 0$, and there exist positive constants c_1 and c_2 such that

$$-g^T s \geq c_1 \|Z^T g\|^2 \quad \textit{and} \quad \|Z^T g\| \geq c_2 \|s\|.$$

Proof. Since the matrix K_{11} has its smallest singular value greater than a positive constant, the compactness of $S(x_0)$ and the boundedness of $\|L^{-1}\|$ guarantee that $\|B_2\|$ is bounded by a constant. Consequently, $\bar{B}_2$ is a bounded block-diagonal matrix whose smallest eigenvalue is greater than c_{TOL}. Consequently, since $\text{In}(B_1) = (n_1, m, 0)$, and the norm of L is bounded, the matrix $L\bar{B}L^T$ has inertia $(n, m, 0)$ and all singular values bounded away from zero by a constant. Since all rows of A are contained in K_{11}, the modification of B_2 only alters H. Consequently,

$$L\bar{B}L^T = \Pi^T \begin{pmatrix} \bar{H} & A^T \\ A & 0 \end{pmatrix} \Pi,$$

where $\bar{H}$ is a modification of H, and z satisfies the equation

$$\begin{pmatrix} \bar{H} & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} s \\ \mu \end{pmatrix} = \begin{pmatrix} -g \\ 0 \end{pmatrix},$$

from which it follows that $As = 0$, and s may be written as

$$s = -Z(Z^T \bar{H} Z)^{-1} Z^T g. \tag{4.16}$$

Since $L\bar{B}L^T$ has inertia $(n, m, 0)$, Lemma 2.1 implies that $Z^T\bar{H}Z$ is positive definite. The boundedness of $\|L\bar{B}L^T\|$ guarantees that $\|Z^T\bar{H}Z\|$ is bounded. Consequently, premultiplication by g^T in (4.16) yields the existence of a positive constant c_1, such that $-g^T s \geq c_1 \|Z^T g\|^2$. Lemmas 2.1 and 2.2 imply that $Z^T\bar{H}Z$ has its smallest eigenvalue greater than a positive constant. Consequently, utilizing norm properties and the boundedness of $\|(Z^T\bar{H}Z)^{-1}\|$, (4.16) guarantees the existence of a positive constant c_2, such that $c_2\|s\| \leq \|Z^T g\|$. □

Note that the direction s computed in Theorem 4.6 is the Newton direction whenever B_2 is sufficiently positive definite, i.e., when the reduced Hessian is sufficiently positive definite. Only if the reduced Hessian is not sufficiently positive definite is the matrix $\bar{B}$ different from B.

If K_{11} is nonsingular, contains all rows of A, and has m negative eigenvalues, but K has more than m negative eigenvalues, there exist feasible directions of negative curvature. Moreover, since $\text{In}(K) = \text{In}(\Pi_1^T K \Pi_1) + \text{In}(\Pi^T K \Pi / \Pi_1^T K \Pi_1)$, it must hold that $\Pi^T K \Pi / \Pi_1^T K \Pi_1$ has at least one negative eigenvalue, and consequently, B_2 must have at least one negative eigenvalue. Because of the structure of B_2, the smallest eigenvalue of B_2 and its corresponding eigenvector are readily available. The following theorem shows that the factors from the LBL^T-factorization of K given in Algorithm 4.1 enable the computation of a feasible direction of negative curvature.

THEOREM 4.7. *Let K be any KKT-matrix in $\mathcal{K}$ with more than m negative eigenvalues. Assume that the LBL^T-factorization of K given in Algorithm* 4.1 *is known. Define $w \equiv \Pi\tilde{w}$, where $\tilde{w}$ satisfies the equation*

$$\begin{pmatrix} L_{11}^T & L_{21}^T \\ 0 & L_{22}^T \end{pmatrix} \begin{pmatrix} \tilde{w}_1 \\ \tilde{w}_2 \end{pmatrix} = \pm\sqrt{-\lambda_{\min}(B_2)} \begin{pmatrix} 0 \\ u_\lambda \end{pmatrix}, \tag{4.17}$$

u_λ is an eigenvector of unit norm corresponding to $\lambda_{\min}(B_2)$, and the sign is chosen so that $g^T d \le 0$. Let d denote the vector composed of the first n elements of w. Then $Ad = 0$ and there exist positive constants c_1 and c_2 such that

$$d^T H d \le -c_1 \lambda_{\min}^2(Z^T H Z) \quad \textit{and} \quad d^T d \le -c_2 \lambda_{\min}(Z^T H Z).$$

Proof. From (4.17) and the definition of w we obtain

$$Kw = \pm\sqrt{-\lambda_{\min}(B_2)}\,\Pi \begin{pmatrix} 0 \\ L_{22} B_2 u_\lambda \end{pmatrix}. \tag{4.18}$$

Since K_{11} contains all rows of A, (4.18) yields $Ad = 0$. It follows from (4.17) that

$$d^T H d = w^T K w = -\lambda_{\min}^2(B_2). \tag{4.19}$$

Using the boundedness of $\|L\|$ and the identity $L_{22} B_2 L_{22}^T = \Pi^T K \Pi / \Pi_1^T K \Pi_1$, Lemma 3.1 guarantees the existence of a positive constant $\bar{c}_1$ such that

$$\lambda_{\min}(B_2) \le \bar{c}_1 \lambda_{\min}(\Pi^T K \Pi / \Pi_1^T K \Pi_1) < 0. \tag{4.20}$$

Since K_{11} has inertia $(n_1, m, 0)$, it contains a nonsingular $m \times m$ submatrix A_B of A. Without loss of generality, we may assume that the partition of A is such that $A = (A_B \;\; A_N)$, where A_N is the remaining submatrix of A. Let $\bar{Z}$ be a reduced-gradient-type null-space matrix generated by A_B defined as

$$\bar{Z} = \begin{pmatrix} -A_B^{-1} A_N \\ I \end{pmatrix}.$$

Properties of $\Pi^T K \Pi / \Pi_1^T K \Pi_1$ are independent of the pivoting order in which K_{11} is created. Without loss of generality, assume that K_{11} was created by first factorizing the $2m \times 2m$ principal submatrix of K containing A_B. Then, with appropriate substitution in the definition of S in Lemma 4.2, it follows that $\bar{Z}^T H \bar{Z}$ is a permuted version of the Schur complement of this $2m \times 2m$ principal submatrix in $\Pi^T K \Pi$. Moreover, the $2m \times 2m$ principal submatrix has inertia $(m, m, 0)$, and since the inertia of $(\Pi^T K \Pi / \Pi_1^T K \Pi_1)$ is $(n_1, m, 0)$, it must hold that when the $2m \times 2m$ principal submatrix has been processed, the remaining $n_1 - m$ pivots until K_{11} has been processed are all positive. Hence, [FGM89B, Lemma 3.3] implies that

$$\lambda_{\min}(\Pi^T K \Pi / \Pi_1^T K \Pi_1) \le \lambda_{\min}(\bar{Z}^T H \bar{Z}) < 0, \tag{4.21}$$

where the second inequality follows from Lemma 2.1. Since there exist only a finite number of nonsingular $m \times m$ submatrices of A, it may be asserted that there exists a positive constant $\bar{c}_2$ such that

$$\lambda_{\min}(\bar{Z}^T H \bar{Z}) \le \bar{c}_2 \lambda_{\min}(Z^T H Z) < 0. \tag{4.22}$$

Combining (4.19), (4.20), (4.21), and (4.22), there exists a positive constant c_1 such that $d^T H d \le -c_1 \lambda_{\min}^2(Z^T H Z)$.

It remains to show the existence of c_2. From (4.17) and the definition of $\tilde{w}$ we get

$$\|d\| \le \|\tilde{w}\| \le \|L^{-1}\| \sqrt{-\lambda_{\min}(B_2)}. \tag{4.23}$$

Since L is unit lower-triangular with bounded norm, $\|L^{-1}\|$ is bounded. Since $Ad = 0$, $d = Z\bar{d}$ for some $\bar{d}$. Hence, since Z is a fixed matrix with full column rank, we conclude that there is a positive constant $\bar{c}_3$ such that $-d^THd \le -\bar{c}_3\|d\|^2\lambda_{\min}(Z^THZ)$. The existence of c_2 now follows by combining (4.19) and (4.23) and using the boundedness of $\|L^{-1}\|$ and the existence of $\bar{c}_3$. □

The strategy for choosing pivots given in this section allows the construction of a nonsingular principal submatrix of K that contains all rows of A and has exactly m negative eigenvalues. This strategy only permits pivots of types H^+, A^-, and HA until all rows of A have been processed. However, as the following lemma shows, if the *full* Hessian is sufficiently positive definite, the pivot strategy described in this section is of no impact at all, since *all* pivots in such circumstances are of types H^+, A^-, and HA.

LEMMA 4.8. *If H is positive definite and A has full row rank, then the only pivots that occur in a regular LBL^T-factorization of K are of type H^+, A^-, or HA.*

Proof. Since K is nonsingular, no pivots of type H^0 or A^0 are used. The positive definiteness of H implies that any principal submatrix of H is positive definite. Consequently, *any* nonsingular principal submatrix K_{11} has inertia $(n_1, m_1, 0)$. This property cannot hold if pivots of type H^-, A^+, HH, or AA are used. □

The conclusion of the lemma does not hold if the Hessian is not positive definite. Consequently, the strategy described in this section may require permutations in addition to those required for a regular LBL^T-algorithm. For example, a pivot of type H^- would not be accepted unless all rows of A have been processed. From the point of view of sparsity, these additional permutations are undesirable since we wish to minimize the change to the ordering determined by the analyze phase.

One possible scheme for choosing the pivots of types H^+, A^-, and HA is to use m pivots of type HA. Such a scheme has been proposed by Gill, Murray, Saunders, and Wright [GMSW87B] in the context of large-scale quadratic programming. They name the HA pivot a *tile* and refer to this scheme as tiling. A scheme of this type automatically fulfills the inertia requirements of Algorithm 4.1. For specific applications, it is known how to choose m tiles as the initial set of pivots. However, how to choose such pivots efficiently for general matrices is not known.

Finally, if the problem is unconstrained, the matrix K equals H, and the pivot strategy described in this section equals the strategy of a regular LBL^T-factorization, since if A does not exist, all rows of A are processed at the very first step of the factorization. The computed directions are then equivalent to those computed by Moré and Sorensen [MS79].

5. Necessary pivot conditions.

In the previous section it was shown that a sufficient condition for the computation of a descent direction and a direction of negative curvature is to use pivots of types H^+, A^-, and HA until all rows of A have been processed. This scheme creates a principal submatrix $\Pi_1^TK\Pi_1$ that contains all rows of A, has inertia $(n_1, m, 0)$, and has a smallest singular value greater than a positive constant. However, it may also be necessary to obtain such a principal submatrix in order to compute a feasible descent direction and a feasible direction of negative curvature, as in the scheme of Moré and Sorensen [MS79]. In the example of §3.4, since $\Pi = I$, the only principal submatrix obtained in the factorization that contains all rows of A is K itself. If a block of type A^-, AA, or A^0 is modified, then the zero-part of K is altered and the subsequent search direction is not feasible. Hence, in the example of §3.4, the only part of B that may be modified is the second element. However, it was shown that such a modification altered both the A-part and the zero-part of K. Similarly, when computing the direction of negative curvature using an eigenvector of B as a right-hand side vector, only eigenvectors corresponding to blocks of types H^- and HH may be used. Again, the example of §3.4 shows that solving for the eigenvector corresponding to the second element of B does not yield a feasible descent direction.

It was shown in §4 that a sufficient condition for obtaining a nonsingular principal submatrix containing all rows of A with m negative eigenvalues was to use pivots of types H^+, A^-, and HA until all rows of A had been processed. Such a scheme may appear unduly restrictive. However, in this section it is shown that these conditions are also necessary to guarantee obtaining such a principal submatrix in a single factorization. We utilize the notation of §4 and let $\Pi_1^T K \Pi_1$ denote a leading principal submatrix of $\Pi^T K \Pi$, for which the factors L_{11} and B_1 are known. Also, to simplify the notation, the related matrix K_{11} from (4.2) is used.

The following lemma shows that unless the inertia of $\Pi_1^T K \Pi_1$ is kept equal to $(n_1, m_1, 0)$ until all rows of A have been processed, it may happen that no nonsingular principal submatrix with m negative eigenvalues, containing K_{11} and all rows of A, exists. Consequently, the results from §4 are not applicable and there is no guarantee that we are able to compute descent directions and directions of negative curvature, as described in §4.

LEMMA 5.1. *Assume that $n > m$. If $\operatorname{In}(K_{11}) = (n_1 - t, m_1 + t, 0)$, where $t > 0$, the matrix A may be such that any nonsingular principal submatrix of K, containing K_{11} and all rows of A, has more than m negative eigenvalues, independently of H_{12}, H_{21}, and H_{22}.*

Proof. If $t > 0$, Lemma 2.1 implies that there exists a nonzero vector p_1 such that $p_1^T H_{11} p_1 < 0$ and $A_{11} p_1 = 0$. If $n > m$, and A has full row rank, the nullspace of A has dimension $n - m$, which is at least one. Consequently, A may be such that $A_{21} p_1 = 0$, in which case

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} p_1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\begin{pmatrix} p_1^T & 0 \end{pmatrix} \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \begin{pmatrix} p_1 \\ 0 \end{pmatrix} = p_1^T H_{11} p_1 < 0.$$

Consequently, Lemma 2.1 implies that K has more than m negative eigenvalues. If *any* nonsingular principal submatrix of K containing K_{11} and all rows of A is considered, by extending the vector p_1 with the appropriate number of zero elements, it follows that such a principal submatrix has more than m negative eigenvalues. This property depends only on the matrices A and K_{11}, and therefore it is independent of the matrices H_{12}, H_{21}, and H_{22}. □

The following example illustrates that even when the reduced Hessian is positive semidefinite, if it is singular, then the restrictions on the choice of pivots given in Algorithm 4.1 are necessary to ensure that a nonsingular K_{11} exists with the number of negative eigenvalues equal to m_1. Let

$$H = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} I & 0 \end{pmatrix}, \tag{5.1}$$

for which the reduced Hessian is zero. If the corresponding K-matrix is factorized, the first pivot must be either of type HA or HH. If an element of type HH is chosen as first pivot, it is not possible to obtain a nonsingular principal submatrix of K, containing this pivot and having as many negative eigenvalues as it contains rows of A. Consequently, Theorem 4.6 does not apply. For this situation, when the reduced Hessian is positive semidefinite and singular, Conn and Gould [CG84] have shown how to compute a descent direction from a single factorization of K. That scheme is not considered here, since it has not been shown that the descent direction given by that scheme satisfies the conditions of Theorem 4.6. In practice, however, the scheme of Conn and Gould [CG84] may be a viable alternative.

On the other hand, it is of interest to detect as early as possible whether K has inertia different from $(n, m, 0)$. However, the next lemma shows that given *any* nonsingular principal submatrix K_{11} with no more than m negative eigenvalues, it may hold that K is nonsingular and has m negative eigenvalues.

LEMMA 5.2. *If* $\mathrm{In}\,(K_{11}) = (n_1 - t, m_1 + t, 0)$, *where* $m_1 + t \leq m$, *it may hold that* $\mathrm{In}\,(K) = (n, m, 0)$.

Proof. If $\mathrm{In}\,(K_{11}) = (n_1 - t, m_1 + t, 0)$, where $m_1 + t \leq m$, it may hold that K_{11} can be expanded with all rows of H so that $\mathrm{In}\,(K_{11}) = (n - \bar{t}, m_1 + \bar{t}, 0)$, where $m_1 + t \leq m_1 + \bar{t} \leq m$. Consequently, if we let $\bar{Z}$ denote a matrix whose columns form an orthonormal basis for the null space of the rows of A contained in the expanded K_{11}, Lemma 2.1 yields $\mathrm{In}\,(\bar{Z}^T H \bar{Z}) = (n - m_1 - \bar{t}, \bar{t}, 0)$. Let v_i, $i = 1, \ldots, n - m_1$, denote orthonormal eigenvectors corresponding to eigenvalues λ_i of $\bar{Z}^T H \bar{Z}$, where the eigenvectors are sorted so that $\lambda_i \leq \lambda_{i+1}$. If the remaining rows of A equal $v_i^T \bar{Z}^T$, $i = 1, \ldots, m - m_1$, then A has full row rank and $p^T H p > 0$ for all $p \neq 0$ such that $Ap = 0$. Consequently, Lemma 2.1 yields $\mathrm{In}\,(K) = (n, m, 0)$. □

6. A sufficient pivoting method.

In this section, we discuss algorithms that generate a sequence $\{x_k\}_{k=0}^{\infty}$ to solve (1.1) that use the descent directions and directions of negative curvature described in Theorems 4.6 and 4.7. To complete the description of an algorithm, we need to define precisely how s_k and d_k are utilized to generate x_{k+1} from x_k. In all cases discussed, x_{k+1} is formed from x_k by adding a linear combination of s_k and d_k. Consequently, since $As_k = 0$ and $Ad_k = 0$, an iterate x_k is of the form $x_k = x_0 + Z\bar{x}_k$. It follows that the sequence $\{\bar{x}_k\}_{k=0}^{\infty}$ may be viewed as a sequence that attempts to find a minimizer to the unconstrained problem

$$\text{(6.1)} \qquad \underset{\bar{x} \in \Re^{n-m}}{\text{minimize}} \quad \bar{f}(\bar{x}) = f(x_0 + Z\bar{x}).$$

This means we may use known convergence results for unconstrained optimization.

6.1. Linesearch conditions.

For unconstrained minimization, Moré and Sorensen [MS79] discuss various methods for combining a descent direction $\bar{s}_k$ and a direction of negative curvature $\bar{d}_k$ to form an iterative sequence with desirable convergence properties. In this section, we review these conditions and discuss how they may be used with either a regular linesearch or a curvilinear linesearch. To fit with the discussion of this paper, we consider minimizing the problem (6.1) and let $\{\bar{x}_k\}_{k=0}^{\infty}$ denote the generated iterative sequence, $\bar{s}_k$ denote the descent direction, and $\bar{d}_k$ denote the direction of negative curvature.

First, we introduce a more generic notation, where we do not specify how $\bar{s}_k$ and $\bar{d}_k$ are combined. At each iterate $\bar{x}_k$, define the univariate function $\phi_k(\alpha)$ as

$$\text{(6.2)} \qquad \phi_k(\alpha) = \bar{f}(\bar{x}_k + \Gamma_k(\alpha))$$

for a twice-continuously differentiable function $\Gamma_k : \Re \to \Re^{n-m}$, where it is required that $\phi_k(0) = \bar{f}(\bar{x}_k)$ and $\phi'_k(0) \leq 0$. Let $\mu \in (0, \frac{1}{2})$ and $\eta \in [\mu, 1)$ denote preassigned constants. The next iterate $\bar{x}_{k+1}$ is defined as $\bar{x}_{k+1} = \bar{x}_k + \Gamma_k(\alpha_k)$ for a positive scalar $\alpha_k \in (0, \alpha_{\max}]$, for which we require the following sufficient-decrease conditions. (If $\phi'_k(0) = 0$ and $\phi''_k(0) \geq 0$, the algorithm terminates, and we may define $\alpha_k = 0$.)

If $\phi''_k(0) \geq 0$, α_k is determined such that either $\alpha_k \in [0, \alpha_{\max}]$ satisfies

$$\phi_k(\alpha_k) \leq \phi_k(0) + \mu\phi'_k(0)\alpha_k \quad \text{and} \quad |\phi'_k(\alpha_k)| \leq \eta|\phi'_k(0)| \tag{6.3}$$

or

$$\phi_k(\alpha_k) \leq \phi_k(0) + \mu\phi'_k(0)\alpha_k \quad \text{and} \quad \alpha_k = \alpha_{\max}. \tag{6.4}$$

If $\phi''_k(0) < 0$, then α_k is required to satisfy either $\alpha_k \in (0, \alpha_{\max}]$ and

$$\phi_k(\alpha_k) \leq \phi_k(0) + \mu(\phi'_k(0)\alpha_k + \tfrac{1}{2}\phi''_k(0)\alpha_k^2) \tag{6.5a}$$

and

$$|\phi'_k(\alpha_k)| \leq \eta|\phi'_k(0) + \phi''_k(0)\alpha_k| \tag{6.5b}$$

or

$$\phi_k(\alpha_k) \leq \phi_k(0) + \mu(\phi'_k(0)\alpha_k + \tfrac{1}{2}\phi''_k(0)\alpha_k^2) \quad \text{and} \quad \alpha_k = \alpha_{\max}. \tag{6.6}$$

For the case $\phi''_k(0) \geq 0$, conditions (6.3) and (6.4) are steplength rules that have been discussed by Gill, Murray, Saunders, and Wright [GMSW79]. For the case $\phi''_k(0) < 0$, conditions similar to (6.5) and (6.6) have been proposed by Moré and Sorensen [MS79], the only difference being that in Moré and Sorensen [MS79], condition (6.5b) is stated without absolute value. However, it follows from the proof of Lemma 5.2 in Moré and Sorensen [MS79] that α_k is well defined.

Combining the results from Gill, Murray, Saunders, and Wright [GMSW79] and Moré and Sorensen [MS79], we arrive at the following lemma.

LEMMA 6.1. *Consider an iterative sequence* $\{\bar{x}_k\}_{k=0}^{\infty}$ *generated from* (6.2) *and* (6.3)–(6.6). *If the level set* $\bar{S}(\bar{x}_0) = \{\bar{x} : \bar{f}(\bar{x}) \leq \bar{f}(\bar{x}_0)\}$ *is compact,* $|\phi'_k(0)|$ *is bounded, and* $|\phi''_k(0)|$ *is bounded, then*

$$\lim_{k\to\infty} \phi'_k(0) = 0 \quad \textit{and} \quad \liminf_{k\to\infty} \phi''_k(0) \geq 0.$$

Proof. It suffices to ensure that the statement of the lemma holds for any subsequence of a generated sequence $\{\bar{x}_k\}_{k=0}^{\infty}$. For a subsequence $\{x_k\}_{k\in J_1}$, where $\phi''_k(0) \geq 0$, it follows from Gill, Murray, Saunders, and Wright [GMSW79] that $\lim_{k\in J_1, k\to\infty} \phi'_k(0) = 0$. For a subsequence $\{x_k\}_{k\in J_2}$, where $\phi''_k(0) < 0$, it follows from [MS79, p. 18] that $\lim_{k\in J_2, k\to\infty} \phi'_k(0) = 0$ and $\lim_{k\in J_2, k\to\infty} \phi''_k(0) = 0$. □

The following lemma shows that if $\bar{s}_k$ is a direction of *sufficient* descent and $\bar{d}_k$ is a direction of *sufficient* negative curvature in the sense required by [MS79, Thm. 6.2], the second-order necessary optimality conditions hold at all limit points.

LEMMA 6.2. *Assume that* $\bar{s}_k$ *has bounded norm and satisfies* $\nabla\bar{f}(\bar{x}_k)^T\bar{s}_k \leq 0$, *and that*

$$\nabla\bar{f}(\bar{x}_k)^T\bar{s}_k \to 0 \quad \textit{implies} \quad \nabla\bar{f}(\bar{x}_k) \to 0 \quad \textit{and} \quad \bar{s}_k \to 0.$$

Also assume that $\bar{d}_k$ has bounded norm, satisfies $\bar{d}_k^T \nabla^2 \bar{f}(\bar{x}_k)\bar{d}_k \le 0$ and $\nabla \bar{f}(\bar{x}_k)^T \bar{d}_k \le 0$, and that

$$\bar{d}_k^T \nabla^2 \bar{f}(\bar{x}_k)\bar{d}_k \to 0 \quad \textit{implies} \quad \lambda_k \to 0 \quad \textit{and} \quad \bar{d}_k \to 0,$$

where λ_k is the minimum of zero and $\lambda_{\min}(\nabla^2 \bar{f}(\bar{x}_k))$.

If the level set $\bar{S}(\bar{x}_0) = \{\bar{x} : \bar{f}(\bar{x}) \le \bar{f}(\bar{x}_0)\}$ is compact, any limit point $\bar{x}^*$ of an iterative sequence $\{\bar{x}_k\}_{k=0}^{\infty}$ generated from (6.3)–(6.6), *using $\phi_k(\alpha) = \bar{f}(\bar{x}_k + \alpha \bar{s}_k + \alpha \bar{d}_k)$ or $\phi_k(\alpha) = \bar{f}(\bar{x}_k + \alpha^2 \bar{s}_k + \alpha \bar{d}_k)$, has the properties*

$$\nabla \bar{f}(\bar{x}^*) = 0 \quad \textit{and} \quad \lambda_{\min}(\nabla^2 \bar{f}(\bar{x}^*)) \ge 0.$$

Proof. To make the notation less cumbersome, it may without loss of generality be assumed that $\bar{x}_k$ converges to $\bar{x}^*$.

If $\Gamma_k(\alpha) = \alpha \bar{s}_k + \alpha \bar{d}_k$ we obtain $\phi_k'(0) = (\bar{s}_k + \bar{d}_k)^T \nabla \bar{f}(\bar{x}_k)$ and $\phi_k''(0) = (\bar{s}_k + \bar{d}_k)^T \nabla^2 \bar{f}(\bar{x}_k)(\bar{s}_k + \bar{d}_k)$. Utilizing $\nabla \bar{f}(\bar{x}_k)^T \bar{d}_k \le 0$, it follows that $\phi_k'(0) \le \bar{s}_k^T \nabla \bar{f}(\bar{x}_k) \le 0$. Consequently, the assumed properties of $\bar{s}_k$ and Lemma 6.1 now yield $\bar{s}_k \to 0$ and $\nabla \bar{f}(\bar{x}^*) = 0$. Now, since $\bar{s}_k \to 0$ and $\|\bar{d}_k\|$ and $\|\nabla^2 \bar{f}(\bar{x}_k)\|$ are bounded, $\phi_k''(0) \to 0$ implies $\bar{d}_k^T \nabla^2 \bar{f}(\bar{x}_k)\bar{d}_k \to 0$. Consequently, the assumed properties of $\bar{d}_k$ and Lemma 6.1 now imply $\bar{d}_k \to 0$ and $\lambda_{\min}(\nabla^2 \bar{f}(\bar{x}^*)) \ge 0$.

If $\Gamma_k(\alpha) = \alpha^2 \bar{s}_k + \alpha \bar{d}_k$, we obtain $\phi_k''(0) = 2\bar{s}_k^T \nabla \bar{f}(\bar{x}_k) + \bar{d}_k^T \nabla^2 \bar{f}(\bar{x}_k)\bar{d}_k$. Again, Lemma 6.1 and the assumed properties of $\bar{s}_k$ and $\bar{d}_k$ now yield the desired result. See Moré and Sorensen [MS79] for details. □

6.2. Convergence properties. In this section, it is demonstrated that the descent direction s_k and the direction of negative curvature d_k from Theorems 4.6 and 4.7 are such that the convergence analysis reviewed in the previous section may be applied directly, upon observing that $s_k = Z\bar{s}_k$ and $d_k = Z\bar{d}_k$.

THEOREM 6.3. *Consider a sequence $\{x_k\}_{k=0}^{\infty}$, generated from* (6.3)–(6.6), *using $\phi_k(\alpha) = f(x_k + \alpha s_k + \alpha d_k)$ or $\phi_k(\alpha) = f(x_k + \alpha^2 s_k + \alpha d_k)$, where s_k is from Theorem* 4.6 *and d_k is from Theorem 4.7 when $Z^T \nabla^2 f(x_k) Z$ is not positive semidefinite; otherwise $d_k = 0$. Any limit point x^* of such a sequence has the properties*

$$Z^T \nabla f(x^*) = 0 \quad \textit{and} \quad \lambda_{\min}(Z^T \nabla^2 f(x^*) Z) \ge 0.$$

Proof. First note that $\nabla \bar{f}(\bar{x}_k) = Z^T \nabla f(x_k)$ and $\nabla^2 \bar{f}(\bar{x}_k) = Z^T \nabla^2 f(x_k) Z$. Consequently, it suffices to show that if s_k is defined from Theorem 4.6 and d_k is defined from Theorem 4.7, they satisfy the sufficiency conditions of Lemma 6.2 when written as $s_k = Z\bar{s}_k$ and $d_k = Z\bar{d}_k$, respectively.

Theorem 4.6 implies that s_k has bounded norm and satisfies $\nabla f(x_k)^T s_k \le 0$, and that $\nabla f(x_k)^T s_k \to 0$ implies $Z^T \nabla f(x_k) \to 0$ and $s_k \to 0$. Since Z is a fixed matrix with full column rank, $s_k \to 0$ implies $\bar{s}_k \to 0$.

Theorem 4.7 implies that d_k has bounded norm, satisfies $d_k^T \nabla^2 f(x_k) d_k \le 0$ and $\nabla f(x_k)^T d_k \le 0$, and that $d_k^T \nabla^2 f(x_k) d_k \to 0$ implies

$$\min\{\lambda_{\min}(Z^T \nabla^2 f(x_k) Z), 0\} \to 0 \quad \text{and} \quad d_k \to 0.$$

Since Z is a fixed matrix with full column rank, $d_k \to 0$ implies $\bar{d}_k \to 0$. □

Note that Theorem 6.3 still holds if instead of s_k and d_k we use $\gamma_k s_k$ and $\beta_k d_k$, where $\epsilon_l \le \gamma_k$ and $\beta_k \le \epsilon_u$ for some positive constants ϵ_l and ϵ_u. The best choice of scaling of these directions has yet to be ascertained.

If $\{x_k\}_{k=0}^{\infty}$ converges to a point $\hat{x}^*$ at which the reduced Hessian is positive definite, then $d_k = 0$ for k sufficiently large. Moreover, linesearch conditions (6.3) are satisfied for $\alpha_k = 1$, using $\phi_k(\alpha) = f(x_k + \alpha s_k)$ or $\phi_k(\alpha) = f(x_k + \alpha^2 s_k)$, for k sufficiently large. Theorem 4.6 implies that s_k is the Newton direction whenever the reduced Hessian at x_k is sufficiently positive definite. This implies that the iterates are generated using Newton's method for k sufficiently large, and the asymptotic rate of convergence is quadratic provided $Z^T \nabla^2 f(x) Z$ is Lipschitz continuous in a neighborhood of $\hat{x}^*$ (see Moré and Sorensen [MS84]).

7. Artificial constraints. In §4, we discussed what conditions may be imposed upon the pivots in order to ensure the ability to compute descent directions and directions of negative curvature from a single factorization of K. An alternative strategy for yielding a descent direction or a direction of negative curvature would be to factorize K using a regular LBL^T-factorization, and from the inertia of B, deduce the inertia of the reduced Hessian. If the reduced Hessian has at least one negative eigenvalue, an *artificial constraint* may be added to A, so that the number of positive eigenvalues of K is increased by one, and consequently the number of negative eigenvalues of the reduced Hessian is reduced by one. An artificial constraint is a temporary additional constraint that is not specified in the original problem. The only requirement for an artificial constraint is linear independence from the original constraints and other artificial constraints. Artificial constraints do not restrict the feasible region, since they are only introduced at a particular iterate and may be removed from the problem at any iterate. If a scheme for adding artificial constraints is known, a positive definite reduced Hessian could be obtained by adding a sufficient number of such constraints. Unless the dimension of the reduced Hessian has been reduced to zero, a descent direction can be obtained. A method for computing a direction of negative curvature for a positive definite reduced Hessian in the presence of artificial constraints has been proposed by Forsgren, Gill, and Murray [FGM89A].

However, as the following lemma shows, to find an artificial constraint that reduces the number of negative eigenvalues of the reduced Hessian by one is equivalent to finding a direction of negative curvature in the null space of A. Consider the case when a nonsingular KKT-matrix

$$K = \begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix}$$

is given, where K has more than m negative eigenvalues. The question of finding an additional artificial constraint a such that

$$\bar{K} = \begin{pmatrix} H & A^T & a \\ A & 0 & 0 \\ a^T & 0 & 0 \end{pmatrix}$$

has one more positive eigenvalue than K is equivalent to finding a direction of negative curvature for the reduced Hessian corresponding to K. The precise statement is given in the following lemma and uses the solution of the equation

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ \mu \end{pmatrix} = \begin{pmatrix} a \\ 0 \end{pmatrix}. \tag{7.1}$$

LEMMA 7.1. *If* $\operatorname{In}(\bar{K}) = \operatorname{In}(K) + (1,0,0)$, *then* p *from* (7.1) *is a direction such that* $Ap = 0$ *and* $p^T H p < 0$. *Conversely, if* q *is a direction such that* $Aq = 0$ *and* $q^T H q < 0$, *then* $\operatorname{In}(\bar{K}) = \operatorname{In}(K) + (1,0,0)$ *for* $a = Hq$.

Proof. Assume that $\operatorname{In}(\bar{K}) = \operatorname{In}(K) + (1,0,0)$. Let $w^T = (a^T \;\; 0)$ and let $u^T = (p^T \;\; \mu^T)$. It follows that

$$\bar{K} = \begin{pmatrix} K & w \\ w^T & 0 \end{pmatrix},$$

and that u solves the equation $Ku = w$. Sylvester's law of inertia implies that $w^T K^{-1} w < 0$. Using the identity $Ku = w$, it follows that $u^T K u < 0$. Consequently, (7.1) yields $p^T H p = u^T K u < 0$, and $Ap = 0$.

Assume that q is a vector such that $Aq = 0$ and $q^T H q < 0$. Let $u^T = (q^T \;\; 0)$ and $w = Ku$. We get $w^T K^{-1} w = q^T H q < 0$. If $a = Hq$, it follows that $w^T = (a^T \;\; 0)$. Sylvester's law of inertia implies that $\operatorname{In}(\bar{K}) = \operatorname{In}(K) + (1,0,0)$. □

Consequently, the ability to add an artificial constraint is linked with the ability to compute a direction of negative curvature in the null space of A. As an example, consider the case when no constraints exist, and $H = I - ee^T$, where e is an n-vector with all components one. This matrix has one negative eigenvalue and $n - 1$ positive eigenvalues. If less than n artificial bounds are added to H, the corresponding K-matrix will have a reduced Hessian that is not positive definite. However, if the single artificial constraint e^T is added, the corresponding reduced Hessian is positive definite. Consequently, although there exist artificial constraints to add, we do not know how to compute them directly. Conn and Gould [CG84] have given a computational scheme for obtaining a direction of negative curvature from the LBL^T-factors of K. However, this scheme requires the solution of a system of equations with an $m \times m$ triangle-like matrix.

8. A descent method.

If no attempt is made to avoid altering the KKT-matrix when the reduced Hessian is positive definite, we may consider an algorithm that yields a descent direction in a single factorization. When forming the factors, pivots corresponding to elements of H are modified, if necessary, so that the principal submatrix factorized at each step is nonsingular and has as many positive eigenvalues as it contains rows of H. This modification corresponds to adding to the diagonal of H, and may be expressed as a positive semidefinite diagonal matrix D of the same dimension as H. If the pivots are modified so that the factorized principal submatrix has its smallest singular value bounded away from zero by a constant, this yields a correction matrix D with bounded norm, such that $Z^T(H + D)Z$ is positive definite and has its smallest eigenvalue bounded away from zero by a constant. Such a correction can be computed in a single factorization and requires only the permutations of a regular LBL^T-algorithm. However, in this method, the correction matrix D may be substantial even if the reduced Hessian is positive definite. Therefore, there is no guarantee that this method has the same rate of convergence as Newton's method. For example, if the original ordering of K is used in the factorization, this method will modify H so that $H + D$ is positive definite. Consequently, if $Z^T H Z$ is positive definite, but H is not, the KKT-matrix is modified unnecessarily. On the other hand, if m pivots of type HA are chosen initially, H will be modified only if the reduced Hessian is not sufficiently positive definite. However, the ordering of the latter case is such that the second-order method described in §6 would not require additional permutations.

Given the modified K-matrix, a descent direction p may be obtained by solving the equation

$$\begin{pmatrix} H+D & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ -\mu \end{pmatrix} = \begin{pmatrix} -g \\ 0 \end{pmatrix}. \tag{8.1}$$

Upon termination of the factorization algorithm described in this section, the LBL^T-factors of the left-hand side matrix of (8.1) are known, and the search direction p may be obtained from these factors. Since $Z^T(H + D)Z$ is positive definite with bounded norm and smallest eigenvalue greater than a positive constant, the search direction p from (8.1) is a sufficient descent direction. If an iterative sequence $\{x_k\}_{k=0}^{\infty}$ is generated as $x_{k+1} = x_k + \alpha_k p_k$, where p_k is defined from (8.1), and if, for $\phi_k(\alpha) = f(x_k + \alpha p_k)$, α_k is chosen to satisfy the sufficient-decrease conditions (6.3)–(6.4) (irrespective of $\phi_k''(0)$), known convergence results show that the reduced gradient is zero at all limit points of a sequence $\{x_k\}_{k=0}^{\infty}$. See, for example, Gill, Murray, Saunders, and Wright [GMSW79].

9. A delta-method. In this section, a method based on a diagonal perturbation of the Hessian is described. For a fixed positive constant δ, a multiple δ of the identity matrix is added to H. Given δ, assume that a search direction p is computed at each iteration from the equation

$$\begin{pmatrix} H+\delta I+D & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} p \\ -\mu \end{pmatrix} = \pm \begin{pmatrix} g \\ 0 \end{pmatrix}, \tag{9.1}$$

where D is a positive semidefinite matrix with bounded norm and the sign of the right-hand side vector is chosen so that $g^T p \leq 0$. The matrix D has all elements zero when the reduced Hessian is sufficiently positive definite. Nonzero elements in D arise for two reasons: first, to ensure that the smallest singular value of the left-hand side matrix of (9.1) is bounded away from zero; second, if having formed an initial factorization there are more than m negative eigenvalues, it may be possible to alter B to reduce the number of negative eigenvalues without impacting the property $Ap = 0$. When A is sparse, the possibility exists that there may be some elements of B that can be modified without losing the property $Ap = 0$. The constant δ may be chosen numerically small, for example, in the order of the square-root of the machine precision. Ideally, the ordering in the factorization of the modified K-matrix, where δI is added to H, may then be kept to whatever is satisfactory for preserving sparsity. However, as was pointed out by the example (5.1), the Schur complement may be singular, and there may be no rows of H left, so we cannot guarantee solving equation (9.1) with a single factorization. However, if it is determined that a single factorization is impossible, by refactorizing, utilizing the factorization method described in §8, we may *always* guarantee a positive semidefinite correction matrix D, such that the left-hand side matrix of (9.1) is sufficiently removed from a singular matrix. In practice, we may hope that a small value of δ would not impact the rate of convergence compared with Newton's method in the neighborhood of a local minimizer, where the reduced Hessian is positive definite. Since for most problems it is unlikely that the matrix on the left-hand side of (9.1) is singular, it will almost always be possible to form an initial factorization.

The following lemma shows that, unless $Z^T g = 0$, the search direction from (9.1) is a nontrivial descent direction or a nontrivial direction of negative curvature.

LEMMA 9.1. *The vector p from* (9.1) *satisfies $Ap = 0$. It holds that $p^T g \le 0$ and at least one of the following conditions is satisfied:*

$$p^T g \le -\frac{\delta}{2} p^T p, \tag{9.2a}$$

$$p^T H p \le -\frac{\delta}{2} p^T p. \tag{9.2b}$$

Proof. It follows from (9.1) that $Ap = 0$. The sign of p is always chosen so that $p^T g \le 0$. Utilizing (9.1), we obtain

$$p^T H p \le -p^T g - \delta p^T p - p^T D p. \tag{9.3}$$

Assume that (9.2a) does not hold. Since D is positive semidefinite, (9.3) yields $p^T H p \le -\frac{\delta}{2} p^T p$, and consequently, (9.2b) holds. □

If the same linesearch as described in §6.1 is applied, we can show that the reduced gradient is zero at all limit points of a sequence $\{x_k\}_{k=0}^{\infty}$. At the kth iterate, p_k is defined from (9.1), and we define

$$\phi_k(\alpha) = f(x_k + \alpha p_k). \tag{9.4}$$

Using this definition of ϕ_k, the sufficient-decrease conditions (6.3)–(6.6) are applied. The following theorem shows that the reduced gradient vanishes at all limit points generated by such a sequence.

THEOREM 9.2. *Let $\{x_k\}_{k=0}^{\infty}$ be a sequence such that $x_{k+1} = x_k + \alpha_k p_k$, where p_k is defined by* (9.1) *and α_k is chosen to satisfy the linesearch conditions* (6.3)–(6.6), *using $\phi_k(\alpha) = f(x_k + \alpha p_k)$. Furthermore, assume that the nonnegative diagonal matrix D_k is always chosen so that the matrix of* (9.1) *has bounded norm and smallest singular value bounded away from zero. Then, any limit point x^* satisfies $Z^T \nabla f(x^*) = 0$.*

Proof. Assume that for some positive constant ϵ there exists a subsequence $\{x_k\}_{k \in J}$ such that $\|Z^T g_k\| \ge \epsilon$ for all $k \in J$. Consider any $k \in J$. If $\bar{H}_k$ is defined as $\bar{H}_k = H_k + \delta I + D_k$, it follows from (9.1) that $p_k = Z\bar{p}_k$, where $\bar{p}_k$ satisfies

$$Z^T \bar{H}_k Z \bar{p}_k = \pm Z^T g_k. \tag{9.5}$$

Utilizing (9.5), we deduce

$$\|\bar{p}_k\| \ge \frac{1}{\|Z^T \bar{H}_k Z\|} \|Z^T g_k\|. \tag{9.6}$$

By the statement of the theorem, $\|\bar{H}_k\|$ is bounded, which implies that $\|Z^T \bar{H}_k Z\|$ is bounded. Consequently, since Z is a fixed matrix of full column rank, (9.6) implies that there exists a positive constant c_J such that $\|p_k\| \ge c_J$ for all $k \in J$. Thus, for any $k \in J$, at least one of the following conditions is satisfied:

$$p_k^T g_k \le -\frac{\delta}{2} c_J^2, \tag{9.7a}$$

$$p_k^T H_k p_k \le -\frac{\delta}{2} c_J^2. \tag{9.7b}$$

Since the smallest singular value of the matrix of (9.1) is bounded away from zero, Lemma 2.2 implies that $\|(Z^T \bar{H}_k Z)^{-1}\|$ is bounded. Thus (9.5) and the compactness of $S(x_0)$ imply that $\|p_k\|$ is bounded. Upon observing that $\phi_k'(0) = p_k^T g_k$

and $\phi_k''(0) = p_k^T H_k p_k$, (9.7) contradicts Lemma 6.1. Consequently, the subsequence $\{x_k\}_{k\in J}$ such that $\|Z^T g_k\| \geq \epsilon$ for all $k \in J$ does not exist, and we conclude that $\lim_{k\to\infty} Z^T g_k = 0$. □

However, since the search direction is zero whenever $Z^T g = 0$, it follows that no stronger result than convergence to a first-order point is possible with this method. A slightly modified method may be obtained by letting δ be variable, i.e., by defining a value δ_k at the kth iteration. If, at the kth iteration, the initial LBL^T-factorization has more than m negative eigenvalues, then $\delta_{k+1} > \delta_k$, in order to reduce the amount of negative curvature in the null space of A. Otherwise, $\delta_{k+1} < \delta_k$. If $\{x_k\}_{k=0}^{\infty}$ converges to a solution where $Z^T H Z$ is positive definite, then δ_k may be reduced so that $\lim_{k\to\infty} \delta_k = 0$, ensuring that the asymptotic rate of convergence is identical to that of Newton's method.

10. Discussion. The pivot strategy discussed in §4 is more restrictive than a regular LBL^T-factorization, since it only allows pivots of type H^+, A^-, and HA until all rows of A have been processed. Consequently, if an initial ordering is given by the analyze phase, this pivot strategy is likely to change the ordering more than a regular LBL^T-factorization. To attempt to maintain sparsity of the factors, it is desirable to reduce the number of additional pivots required in the numerical phase. In some circumstances, it may be possible to accept pivots of other types than H^+, A^-, and HA. It was shown in §3.4 that modifying a diagonal element of B of type H^- or HH may alter the A-part and the zero-part of K. However, the only altered elements correspond to nonzeros in the outer product created by the corresponding column or columns of L. It is a simple matter to check if these nonzeros of L correspond to rows of A. If they do not, the particular H^- or HH pivot can be accepted, even if the number of negative eigenvalues of K_{11} exceeds m_1 by doing so. When K is sparse, it may be a common event that the nonzeros do not correspond to rows of A once a significant portion of the rows of A have been processed. If the nonzeros do not correspond to rows of A, the reduced Hessian has at least one negative eigenvalue, and we can still obtain a feasible direction of negative curvature.

Another scheme is to accept any pivot a regular LBL^T-factorization would accept, keeping track of restart points, where K_{11} has inertia $(n_1, m_1, 0)$ and is sufficiently nonsingular. If it turns out that K has inertia $(n, m, 0)$, the reduced Hessian is positive definite and the Newton direction is a descent direction. Otherwise, when forming the factorization, let K_{11} denote the part of K that is factorized when it is discovered from the Schur complement that the inertia of K is different from $(n, m, 0)$. If K_{11} contains all rows of A, has inertia $(n_1, m, 0)$ and is sufficiently nonsingular, the results of §4 apply. If not all rows of A have yet been processed, an attempt may be made to find pivots of types H^+ and A^+ in order to form a K_{11} that is sufficiently nonsingular, contains all rows of A, and has m negative eigenvalues. If this attempt is successful, Theorems 4.6 and 4.7 apply, and the desired directions may be computed. If the attempt is not successful, part of K_{11}, from the latest restart point, may be refactorized, imposing the pivoting strategy of §4.

If the number of positive eigenvalues in the reduced Hessian is large compared to the number of negative eigenvalues, we expect the number of pivots of types H^- and HH to be low. Consequently, if the rows of A are processed early in the factorization, there is a high likelihood that these pivots will occur only after all rows of A have been processed. A new version of MA27 (MA47) allows 2×2 pivots in the analyze phase. See Duff et al. [DGRST89]. Moreover, these pivots, which in our case would be HA pivots, are preferred over 1×1 pivots. We expect this scheme to make the difference between the additional permutation requirements of the scheme of §4 and the additional permutations required by a regular LBL^T-factorization smaller, since in many instances the conditions of §4 will be fulfilled automatically.

It may also be observed that the ability to compute a direction of negative curvature is only required if the reduced gradient is small in norm. The methods described in §§8 and 9 may be applied whenever the norm of the reduced gradient is sufficiently positive. Only at points where the reduced gradient has small norm and the reduced Hessian is not positive definite is it necessary to apply the strategy of §4.

As was discussed in §7, to add a suitable artificial constraint is equivalent to generating a feasible direction of negative curvature. It can be shown that an artificial constraint that is linearly independent of the constraints in A may be found by one solve with K utilizing a suitable right-hand side. If this artificial constraint increases the number of positive eigenvalues of K by one, a direction of negative curvature may be computed as described in Lemma 7.1. Although this is not guaranteed to be the case, in practice, it may be a viable strategy.

Acknowledgments. We thank Philip Gill for many helpful discussions and suggestions clarifying the results presented. We also thank Nick Gould and Per Olov Lindberg for their careful reading of the manuscript and a number of helpful comments.

REFERENCES

[BK77] J. R. Bunch and L. Kaufman, *Some stable methods for calculating inertia and solving symmetric linear systems*, Math. Comp., 31 (1977), pp. 163–179.

[BP71] J. R. Bunch and B. N. Parlett, *Direct methods for solving symmetric indefinite systems of linear equations*, SIAM J. Numer. Anal., 8 (1971), pp. 639–655.

[CG84] A. R. Conn and N. I. M. Gould, *On the location of directions of infinite descent for nonlinear programming algorithms*, SIAM J. Numer. Anal., 21 (1984), pp. 1162–1179.

[DGRST89] I. S. Duff, N. I. M. Gould, J. K. Reid, J. A. Scott, and K. Turner, *The factorization of sparse symmetric indefinite matrices*, Tech. Rep. CSS 236, Computer Science and Systems Division, AERE Harwell, Oxford, England, 1989.

[DR82] I. S. Duff and J. K. Reid, *MA27—a set of Fortran subroutines for solving sparse symmetric sets of linear equations*, Tech. Rep. R-10533, Computer Science and Systems Division, AERE Harwell, Oxford, England, 1982.

[DR83] ———, *The multifrontal solution of indefinite sparse symmetric linear equations*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.

[DS89] J. E. Dennis, Jr. and R. B. Schnabel, *A view of unconstrained optimization*, in Handbooks in Operations Research and Management Science, Vol. 1. Optimization, G. L. Nemhauser, A. H. G. Rinnooy Kan, and M. J. Todd, eds., North Holland, Amsterdam, New York, Oxford, and Tokyo, 1989, Chap. 1, pp. 1–72.

[FF77] R. Fletcher and T. L. Freeman, *A modified Newton method for minimization*, J. Optim. Theory Appl., 23 (1977), pp. 357–372.

[FGM89a] A. Forsgren, P. E. Gill, and W. Murray, *On the identification of local minimizers in inertia-controlling methods for quadratic programming*, Report SOL 89-11, Dept. of Operations Research, Stanford Univ., Stanford, CA, 1989; SIAM J. Matrix Anal. Appl., 12 (1991), pp. 730–746.

[FGM89b] ———, *A modified Newton method for unconstrained minimization*, Report SOL 89-12, Dept. of Operations Research, Stanford Univ., Stanford, CA, 1989.

[FM68] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley and Sons, Inc., New York, London, Sydney, and Toronto, 1968.

[GM74] P. E. Gill and W. Murray, *Newton-type methods for unconstrained and linearly constrained optimization*, Math. Programming, 7 (1974), pp. 311–350.

[GMSW79] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *Two steplength algorithms for numerical optimization*, Report SOL 79-25, Dept. of Operations Research, Stanford Univ., Stanford, CA, 1979.

[GMSW85] ———, *Software and its relationship to methods*, in Numerical Optimization 1984, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1985, pp. 139–159.

[GMSW87a] ———, *Maintaining LU-factors of a general sparse matrix*, Linear Algebra Appl., 88/89 (1987), pp. 239–270.

[GMSW87b] ———, *A Schur-complement method for sparse quadratic programming*, Report SOL 87-12, Dept. of Operations Research, Stanford Univ., Stanford, CA, 1987.

[Gol80] D. Goldfarb, *Curvilinear path steplength algorithms for minimization which use directions of negative curvature*, Math. Programming, 18 (1980), pp. 31–40.

[Gou85] N. I. M. Gould, *On practical conditions for the existence and uniqueness of solutions to the general equality quadratic programming problem*, Math. Programming, 32 (1985), pp. 90–99.

[GV89] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd Ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[KD79] S. Kaniel and A. Dax, *A modified Newton's method for unconstrained minimization*, SIAM J. Numer. Anal., 16 (1979), pp. 324–331.

[McC77] G. P. McCormick, *A modification of Armijo's step-size rule for negative curvature*, Math. Programming, 13 (1977), pp. 111–115.

[MP78] H. Mukai and E. Polak, *A second-order method for unconstrained optimization*, J. Optim. Theory Appl., 26 (1978), pp. 501–513.

[MS79] J. J. Moré and D. C. Sorensen, *On the use of directions of negative curvature in a modified Newton method*, Math. Programming, 16 (1979), pp. 1–20.

[MS84] ———, *Newton's method*, in Studies in Mathematics, Vol. 24. Studies in Numerical Analysis, G. H. Golub, ed., The Mathematical Association of America, 1984, pp. 29–82.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 588–597, April 1993

PERTURBATION BOUNDS FOR THE POLAR DECOMPOSITION*

ROY MATHIAS†

Abstract. Let $M_n(F)$ denote the space of matrices over the field F. Given $A \in M_n(F)$ define $|A| \equiv (A^*A)^{1/2}$ and $U(A) \equiv A|A|^{-1}$ assuming A is nonsingular. Let $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_n(A) \geq 0$ denote the ordered singular values of A.

Majorization results are obtained relating the singular values of $U(A+\Delta A) - U(A)$ and those of A and ΔA. In particular, it is shown that if $A, \Delta A \in M_n(R)$ and $\sigma_1(\Delta A) < \sigma_n(A)$, then for any unitarily invariant norm $\|\cdot\|$, $\|U(A+\Delta A) - U(A)\| \leq 2[\sigma_{n-1}(A) + \sigma_n(A)]^{-1}\|\Delta A\|$. Similar results are obtained for matrices with complex entries.

Also considered is the unitary Procrustes problem: $\min\{\|A - UB\| : U \in M_n(C), U^*U = I\}$ where $A, B \in M_n(C)$, and a unitarily invariant norm $\|\cdot\|$ are given. It was conjectured that if U is unitary and U^*BA^* is positive semidefinite then U must be a solution to the unitary Procrustes problem for all unitarily invariant norms. The conjecture is shown to be false.

Key words. polar decomposition, perturbation bound, unitarily invariant norm, majorization, unitary Procrustes problem

AMS(MOS) subject classifications. 15A45, 15A12, 15A48, 15A60

1. Introduction.

Let $M_n(F)$ denote the space of $n \times n$ matrices over the field F. (We will take $F = R$ or $F = C$.) Given $A \in M_n(F)$ there is a unitary $U \in M_n(F)$ and a positive semidefinite $P \in M_n(F)$ such that $A = UP$ (this is the *polar decomposition* of A). It can be seen that $P = (A^*A)^{1/2} \equiv |A|$. When A is nonsingular then so is $|A|$ and hence U must be $A|A|^{-1}$. For a nonsingular A we define $U(A) \equiv A|A|^{-1}$. We say that a norm $\|\cdot\|$ on $M_n(F)$ is *unitarily invariant* if

$$\|A\| = \|UAV\| \quad \text{for all unitary } U, V \in M_n(F).$$

We will say more about these norms later in this section.

In this paper we consider bounds on $\|U(A) - U(B)\|$ in terms of $\|A - B\|$ for unitarily invariant norms $\|\cdot\|$. The perturbation theory of the polar factor in the polar decomposition is of interest, as it is often necessary to compute $U(A)$ numerically. (See [4] for a variety of applications of the polar decomposition. For a more recent application, see [12], where the authors use the polar decomposition in a crucial way to compute block Householder transformations.) The case where $\|\cdot\|$ is the Frobenius norm has been studied by several authors and the best bounds to date are given by Barrlund in [2]. We will strengthen his result slightly, generalize it to all unitarily invariant norms, and deduce the corresponding result about the condition of the function $U(\cdot)$. It will be seen that the cases $F = R$ and $F = C$ are different.

The perturbation theory of the map $A \to |A|$ is also of interest. We will briefly review what is known about it at the end of the paper.

We define the *Hadamard product* of $A = [a_{ij}], B = [b_{ij}] \in M_n(F)$ by

$$A \circ B \equiv [a_{ij}b_{ij}].$$

* Received by the editors July 29, 1991; accepted for publication (in revised form) October 18, 1991.

† Department of Mathematics, College of William & Mary, Williamsburg, Virginia 23187 (na.mathias@na-net.ornl.gov). This author is currently visiting the Institute for Mathematics and Its Applications, University of Minnesota, 514 Vincent Hall, Minneapolis, Minnesota 55455. This research was supported by an Eliezer Naddor Postdoctoral Fellowship in the Mathematical Sciences from the Johns Hopkins University during the year 1989–1990 while the author was in residence at the Department of Computer Science, Stanford University.

The rest of this section is devoted to background information on singular values and unitarily invariant norms. Given $A \in M_n(F)$ we use $\sigma_1(A) \geq \cdots \geq \sigma_n(A)$ to denote its singular values. For each $k = 1, \ldots, n$ we define the *Ky Fan k-norm*, which is unitarily invariant, by

$$|||A|||_k \equiv \sum_{i=1}^{k} \sigma_i(A).$$

Note that $|||A|||_1 = \sigma_1(A)$ is the spectral norm or operator norm of A (often denoted $\|A\|_2$). The norm $|||\cdot|||_n$ on $M_n(F)$ is called the trace norm, and can be represented as $|||A|||_n = \operatorname{tr} |A|$. We will denote the Frobenius norm by $\|\cdot\|_F$, i.e., $\|X\|_F^2 \equiv \operatorname{tr} X^*X$. von Neumann showed (see, e.g., [5, Thm. 7.4.24]) that for every unitarily invariant norm $\|\cdot\|$ on $M_n(F)$ there is a corresponding symmetric gauge function g on R^n (that is, a permutation invariant absolute norm on R^n [6, Def. 3.5.17]) such that

$$\|A\| = g(\sigma_1(A), \ldots, \sigma_n(A)) \quad \text{for all } A \in M_n(F),$$

and that, conversely, every symmetric gauge function $g(\cdot)$ on R^n defines a unitarily invariant norm in this way. The following is a well-known result of Ky Fan [6, Cor. 3.5.9].

LEMMA 1.1. *Let $A, B \in M_n(F)$ be given. Then*

$$\|A\| \leq \|B\| \quad \textit{for all unitarily invariant norms on } M_n(F)$$

if and only if

$$|||A|||_k \leq |||B|||_k, \qquad k = 1, \ldots, n,$$

or equivalently,

$$\sum_{i=1}^{k} \sigma_i(A) \leq \sum_{i=1}^{k} \sigma_i(B), \qquad k = 1, \ldots, n.$$

LEMMA 1.2. *Let $A, B \in M_n(F)$ be given. Then*

$$\sum_{i=1}^{k} [\sigma_{n+1-i}(A) - \sigma_i(B)] \leq \sum_{i=1}^{k} \sigma_{n+1-i}(A+B), \qquad k = 1, \ldots, n. \tag{1.1}$$

Proof. The inequality (1.1) is equivalent to

$$\sum_{i=1}^{k} [\sigma_{n+1-i}(A) - \sigma_{n+1-i}(A+B)] \leq \sum_{i=1}^{k} \sigma_i(B), \qquad k = 1, \ldots, n, \tag{1.2}$$

which we will prove. Take $k \in \{1, \ldots, n\}$. Then

$$\begin{aligned} \sum_{i=1}^{k} [\sigma_{n+1-i}(A) - \sigma_{n+1-i}(A+B)] &\leq \max_{1 \leq i_1 < \cdots i_k \leq n} \sum_{j=1}^{k} |\sigma_{i_j}(A) - \sigma_{i_j}(A+B)| \\ &\leq \sum_{i=1}^{k} \sigma_i(A - (A+B)) \\ &= \sum_{i=1}^{k} \sigma_i(B). \end{aligned}$$

We have used [5, Thm. 7.4.51] for the second inequality. □

2. Results for real matrices. We begin with a basic computational lemma from [2, eq. (2.18a)].

LEMMA 2.1. *Let* $A = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ *and assume that* $\sigma_n > 0$. *Let* $B \in M_n(F)$ *and let* $X(t) = A + tB$. *Then*

$$\left(\frac{d}{dt}U(X(t))|_{t=0}\right)_{ij} = \frac{b_{ij} - \bar{b}_{ji}}{\sigma_i + \sigma_j},$$

or equivalently,

$$\frac{d}{dt}U(X(t))|_{t=0} = C \circ (B - B^*), \tag{2.1}$$

where $C = [(\sigma_i + \sigma_j)^{-1}]$.

The following lemma is essentially the same as [11, Thm. 3.1].

LEMMA 2.2. *Let* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$ *and let* $C \equiv [(\sigma_i + \sigma_j)^{-1}]$. *Let* $B \in M_n(R)$ *be skew-symmetric. Then*

$$|\!|\!|B \circ C|\!|\!|_k = \sum_{i=1}^{k} \sigma_i(B \circ C) \leq \sum_{i=1}^{k} \frac{\sigma_i(B)}{2\hat{\sigma}_i} \leq \frac{1}{2\hat{\sigma}_1} \sum_{i=1}^{k} \sigma_i(B) = \frac{1}{2\hat{\sigma}_1} |\!|\!|B|\!|\!|_k.$$

Using these two results we can prove our first perturbation bound.

THEOREM 2.3. *Let* $A \in M_n(R)$ *be nonsingular. Then for any* $\Delta A \in M_n(R)$ *with* $\sigma_1(\Delta A) < \sigma_n(A)$ *and any unitarily invariant norm* $\|\cdot\|$,

$$\|U(A + \Delta A) - U(A)\| \leq -\frac{2\|\Delta A\|}{|\!|\!|\Delta A|\!|\!|_2} \log\left(1 - \frac{|\!|\!|\Delta A|\!|\!|_2}{\sigma_n(A) + \sigma_{n-1}(A)}\right). \tag{2.2}$$

Proof. By Lemma 1.1 it is sufficient to prove the inequality for the norms $|\!|\!|\cdot|\!|\!|_k$ for $k = 1, \ldots, n$. Let $X(t) = A + t\Delta A$. Then choose $t_0 \in [0, 1]$ and let $X(t_0) = V\Sigma W$ be a singular value decomposition of $X(t_0)$. Then

$$\begin{aligned}
|\!|\!|\tfrac{d}{dt}U(X(t))|_{t=t_0}|\!|\!|_k &= |\!|\!|V \tfrac{d}{dt}U(\Sigma + tV^T\Delta AW^T)|_{t=0}\, W|\!|\!|_k \\
&= |\!|\!|\tfrac{d}{dt}U(\Sigma + tV^T\Delta AW^T)|_{t=0}|\!|\!|_k \\
&= |\!|\!|C \circ [V^T\Delta AW^T - (V^T\Delta AW^T)^T]|\!|\!|_k \\
&\leq \frac{|\!|\!|V^T\Delta AW^T - (V^T\Delta AW^T)^T|\!|\!|_k}{\sigma_n(\Sigma + t_0V^T\Delta AW^T) + \sigma_{n-1}(\Sigma + t_0V^T\Delta AW^T)} \\
&\leq \frac{2|\!|\!|V^T\Delta AW^T|\!|\!|_k}{\sigma_n(\Sigma + t_0V^T\Delta AW^T) + \sigma_{n-1}(\Sigma + t_0V^T\Delta AW^T)} \\
&= \frac{2|\!|\!|\Delta A|\!|\!|_k}{\sigma_n(A + t_0\Delta A) + \sigma_{n-1}(A + t_0\Delta A)} \\
&\leq \frac{2|\!|\!|\Delta A|\!|\!|_k}{\sigma_n(A) + \sigma_{n-1}(A) - t_0|\!|\!|\Delta A|\!|\!|_2}.
\end{aligned}$$

We have used the case $k = 2$ of Lemma 1.2 for the last inequality. Integrating this gives the required bound:

$$|\!|\!|U(A + \Delta A) - U(A)|\!|\!|_k = |\!|\!|\textstyle\int_0^1 \frac{d}{dt}U(X(t))dt|\!|\!|_k$$

$$
\begin{aligned}
&\le \int_0^1 \|\!|\tfrac{d}{dt}U(X(t))|\!\|_k dt\\
&\le \int_0^1 \frac{2|\!|\!|\Delta A|\!|\!|_k}{\sigma_n(A)+\sigma_{n-1}(A)-t|\!|\!|\Delta A|\!|\!|_2}dt\\
&= -\frac{2|\!|\!|\Delta A|\!|\!|_k}{|\!|\!|\Delta A|\!|\!|_2}\log\left(1-\frac{|\!|\!|\Delta A|\!|\!|_2}{\sigma_n(A)+\sigma_{n-1}(A)}\right). \quad \square
\end{aligned}
$$

It is easy to show that our bound when restricted to the Frobenius norm is better than the bound in [2]:

$$\|U(A+\Delta A)-U(A)\|_F \le -\sqrt{2}\log\left(1-\frac{\sqrt{2}\|\Delta A\|_F}{\sigma_n(A)+\sigma_{n-1}(A)}\right).$$

(Just use the fact that $a\log(1-b) > \log(1-ab)$ if $a,b\in(0,1)$ and the fact that $\sqrt{2}\|\Delta A\|_F \ge |\!|\!|\Delta A|\!|\!|_2$.) The reason that our inequality is better than that in [2] is that we used Lemma 1.2 rather than the Wielandt–Hoffman inequality which was used in [2].

Note that if we take $A\in M_n(R)$ with $\sigma_{n-1}(A) > \sigma_n(A)$ then for any unitarily invariant norm $\|\cdot\|$, Theorem 2.3 gives

$$
\begin{aligned}
&\sup\{\|U(A+\Delta A)-U(A)\| : \sigma_1(\Delta A) < \sigma_n(A)\}\\
&\quad\le \sup\left\{-\frac{2\|\Delta A\|}{|\!|\!|\Delta A|\!|\!|_2}\log\left(1-\frac{|\!|\!|\Delta A|\!|\!|_2}{\sigma_n(A)+\sigma_{n-1}(A)}\right) : \sigma_1(\Delta A) < \sigma_n(A)\right\}\\
&\quad\le -n\log\left(1-\frac{2\sigma_n(A)}{\sigma_n(A)+\sigma_{n-1}(A)}\right)\\
&\quad< \infty.
\end{aligned}
$$

In view of this, one might expect that the condition $\sigma_1(\Delta A) < \sigma_n(A)$ in Theorem 2.3 can be replaced by $\sigma_1(\Delta A)+\sigma_2(\Delta A) < \sigma_n(A)+\sigma_{n-1}(A)$. It cannot. Consider, for example,

$$A=\begin{pmatrix}10 & 0\\ 0 & 1\end{pmatrix} \quad\text{and}\quad \Delta A=\begin{pmatrix}0 & 0\\ 0 & -2\end{pmatrix}.$$

It is easy to check that $\sigma_1(\Delta A)+\sigma_2(\Delta A) < \sigma_n(A)+\sigma_{n-1}(A)$ holds for this choice of A and ΔA, but that (2.2) does not.

The following result can be proved in the same way as Theorem 2.3 and can be applied more generally. Neither bound is uniformly better than the other.

THEOREM 2.4. *Let $A, \Delta A\in M_n(R)$ be given. Assume that $A+t\Delta A$ is nonsingular for all $t\in[0,1]$. Then for any unitarily invariant norm $|\!|\!|\cdot|\!|\!|$,*

$$
\begin{aligned}
&|\!|\!|U(A+\Delta A)-U(A)|\!|\!|\\
&\quad\le 2\max\{[\sigma_n(A+t\Delta A)+\sigma_{n-1}(A+t\Delta A)]^{-1} : 0\le t\le 1\}\ |\!|\!|\Delta A|\!|\!|.
\end{aligned}
$$

The following corollary may be deduced from either Theorem 2.3 or Theorem 2.4. The result (2.3) when the norm is the Frobenius norm was first noted in [8, Thm. 2.3].

COROLLARY 2.5. *Let $n \geq 2$ and let $A \in M_n(R)$ be nonsingular. Then the condition of the function $U(\cdot)$ at A with respect to real perturbations and any unitarily invariant norm is*

$$\kappa(A) = 2[\sigma_n(A) + \sigma_{n-1}(A)]^{-1}. \tag{2.3}$$

Proof. That $\kappa(A) \leq 2[\sigma_n(A) + \sigma_{n-1}(A)]^{-1}$ follows from either Theorem 2.3 or Theorem 2.4. Let us show $\kappa(A) \geq 2[\sigma_n(A) + \sigma_{n-1}(A)]^{-1}$. Let A be nonsingular, and let $A = U\Sigma V$ be a singular value decomposition of A. Let

$$B = U\left(0_{n-2} \oplus \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\right) V$$

and let $X(t) = A + tB$. Then Lemma 2.1 gives

$$\frac{d}{dt}U(X(t))|_{t=0} = \frac{2B}{\sigma_n(A) + \sigma_{n-1}(A)}.$$

Thus

$$\begin{aligned}\kappa(A) &\geq \lim_{t\downarrow 0} \frac{\|U(A+tB) - U(A)\|}{\|tB\|} \\ &= \frac{1}{\|B\|} \lim_{t\downarrow 0} \frac{\|U(A+tB) - U(A)\|}{t} \\ &= \frac{1}{\|B\|} \left\| \frac{d}{dt}U(X(t))|_{t=0} \right\| \\ &= \frac{2}{\sigma_n(A) + \sigma_{n-1}(A)}.\end{aligned}$$

If we allow ourselves to bound $|||U(A + \Delta A) - U(A)|||$ in terms of the singular values of ΔA rather than in terms of $|||\Delta A|||$, then we get the next result, which is stronger than Theorem 2.3.

We will need the following notation. Given a finite sequence $\{\sigma_i\}_{i=1}^n$ we will define the sequences $\{\bar{\sigma}_i\}_{i=1}^n$ and $\{\hat{\sigma}_i\}_{i=1}^n$ as follows:

$$\bar{\sigma}_{2i} = \bar{\sigma}_{2i-1} = (\sigma_{2i} + \sigma_{2i-1})/2, \qquad i = 1, \ldots, [n/2]$$

($\bar{\sigma}_n = 0$ if n is odd),

$$\hat{\sigma}_{2i} = \hat{\sigma}_{2i-1} = (\sigma_{n+2-2i} + \sigma_{n+1-2i})/2, \qquad i = 1, \ldots, [n/2]$$

($\hat{\sigma}_n = +\infty$ if n is odd). Notice that if σ_i is an increasing sequence, then $\bar{\sigma}_i$ is increasing while $\hat{\sigma}_i$ is decreasing.

THEOREM 2.6. *Let $A \in M_n(R)$ be nonsingular, let $\Delta A \in M_n(R)$ be such that $\sigma_1(\Delta A) < \sigma_n(A)$, and let $\|\cdot\|$ be any unitarily invariant norm. Then we have the majorization*

$$\sum_{i=1}^k \sigma_i(U(A + \Delta A) - U(A)) \leq -\sum_{i=1}^k \log\left(1 - \frac{\bar{\sigma}_i(\Delta A)}{\hat{\sigma}_i(A)}\right), \quad k = 1, \ldots, n \tag{2.4}$$

and the bound

$$\|U(A + \Delta A) - U(A)\| \leq g\left(\log\left(1 - \frac{\bar{\sigma}_1(\Delta A)}{\hat{\sigma}_1(A)}\right), \ldots, \log\left(1 - \frac{\bar{\sigma}_n(\Delta A)}{\hat{\sigma}_n(A)}\right)\right), \tag{2.5}$$

where g is the symmetric gauge function corresponding to $\|\cdot\|$.

Proof. In view of Lemma 1.1 it is sufficient to prove the majorization (2.4). Let us use the same notation as in the proof of Theorem 2.3. Then, as before, but using the stronger inequality in Lemma 2.2, we have

$$\begin{aligned}\left|\!\left|\!\left| \tfrac{d}{dt}U(X(t))|_{t=t_0} \right|\!\right|\!\right|_k &= |\!|\!| C \circ [V^T\Delta AW^T - (V^T\Delta AW^T)^T] |\!|\!|_k \\ &\leq \sum_{i=1}^{k} \frac{\bar{\sigma}_i(V^T\Delta AW^T - (V^T\Delta AW^T)^T)}{\hat{\sigma}_i(\Sigma + t_0V^T\Delta AW^T)}.\end{aligned} \tag{2.6}$$

The matrix $V^T\Delta AW^T - (V^T\Delta AW^T)^T$ is a real skew-symmetric matrix, so its singular values occur in pairs. This fact, together with the inequality

$$\begin{aligned}\sum_{i=1}^{k} \sigma_i(V^T\Delta AW^T - (V^T\Delta AW^T)^T) &\leq \sum_{i=1}^{k} \sigma_i(V^T\Delta AW^T) \\ &\quad + \sum_{i=1}^{k} \sigma_i((V^T\Delta AW^T)^T) \\ &= 2\sum_{i=1}^{k} \sigma_i(\Delta A),\end{aligned}$$

gives

$$\sum_{i=1}^{k} \sigma_i(V^T\Delta AW^T - (V^T\Delta AW^T)^T) \leq 2\sum_{i=1}^{k} \bar{\sigma}_i(\Delta A), \qquad k = 1, \ldots, n. \tag{2.7}$$

Since $\hat{\sigma}_1(\Sigma + t_0V^T\Delta AW^T) \geq \hat{\sigma}_2(\Sigma + t_0V^T\Delta AW^T) \geq \cdots \geq \hat{\sigma}_n(\Sigma + t_0V^T\Delta AW^T)$ the majorization (2.7) and the inequality (2.6) together give

$$\left|\!\left|\!\left| \tfrac{d}{dt}U(X(t)) \right|\!\right|\!\right|_k \leq 2\sum_{i=1}^{k} \frac{\bar{\sigma}_i(\Delta A)}{2\hat{\sigma}_i(A + t\Delta A)}, \qquad k = 1, \ldots, n. \tag{2.8}$$

Lemma 1.2 implies

$$\sum_{i=1}^{k} \hat{\sigma}_i(A + t\Delta A) \geq \sum_{i=1}^{k} \hat{\sigma}_i(A) - t\bar{\sigma}_i(\Delta A), \qquad k = 1, \ldots, n.$$

Since $\{\hat{\sigma}_i(A + t\Delta A)\}_{i=1}^n$ is a nonnegative increasing sequence, Theorem 5.A.2.iii (with $g(t) = t^{-1}$) in [9] together with the previous majorization gives

$$\sum_{i=1}^{k} \hat{\sigma}_i(A + t\Delta A)^{-1} \leq \sum_{i=1}^{k} [\hat{\sigma}_i(A) - t\bar{\sigma}_i(\Delta A)]^{-1}, \qquad k = 1, \ldots, n.$$

Combining this with the inequality (2.8) we have the bound

$$\left|\!\left|\!\left| \tfrac{d}{dt}U(X(t)) \right|\!\right|\!\right|_k \leq \sum_{i=1}^{k} \frac{\bar{\sigma}_i(\Delta A)}{\hat{\sigma}_i(A) - t\bar{\sigma}_i(\Delta A)}, \qquad k = 1, \ldots, n.$$

Integrating this bound, as in the proof of Theorem 2.3, gives the majorization (2.4). □

In the case of $|||\cdot|||_1$ (the spectral norm) this theorem gives a slight improvement over the bound given by Theorem 2.3. The factor

$$\frac{2|||\Delta A|||_1}{|||\Delta A|||_2} = \frac{2\sigma_1(\Delta A)}{\sigma_1(\Delta A)+\sigma_2(\Delta A)} \geq 1$$

is replaced by 1.

3. Complex matrices.

The formula for the derivative in Lemma 2.1 is still valid when A and ΔA are complex. However, the bound in Lemma 2.2 is not valid for possibly complex skew-symmetric matrices B. For this reason the results in the previous section are not valid for complex matrices. For example, it is easy to check that if

$$A = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \Delta A = \begin{pmatrix} 0 & 0 \\ 0 & i/2 \end{pmatrix},$$

then $\sigma_1(\Delta A) < \sigma_n(A)$, but (2.2) does not hold for the spectral norm.

In the complex case we have the following bound on the norm of a Hadamard product (which is a special case of [1, eq. (1c)]).

LEMMA 3.1. *Let $B, C \in M_n(F)$ be given. Assume that C is positive semidefinite and that its ordered main diagonal entries are $c_1 \geq \cdots \geq c_n$. Then*

$$\sum_{i=1}^{k} \sigma_i(B\circ C) \leq \sum_{i=1}^{k} c_i\sigma_i(B), \qquad k = 1,\ldots,n.$$

Given $\sigma_1 \geq \cdots \geq \sigma_n > 0$, one can show that the Cauchy matrix $C = [(\sigma_i+\sigma_j)^{-1}]$ is positive semidefinite using the idea in [5, Problem 17, p. 401]: For any $x \in C^n$,

$$x^*Cx = \sum_{ij} \bar{x}_i(\sigma_i+\sigma_j)^{-1}x_j = \int_0^\infty \left|\sum_{i=1}^{n} x_i e^{-c_i t}\right|^2 dt \geq 0.$$

Thus, using Lemma 3.1 in place of Lemma 2.2, we have the following analogues of Theorems 2.3, 2.4, and 2.6.

THEOREM 3.2. *Let $A \in M_n(C)$ be nonsingular. Then for any $\Delta A \in M_n(C)$ with $\sigma_1(\Delta A) < \sigma_n(A)$ and any unitarily invariant norm $\|\cdot\|$,*

$$\|U(A+\Delta A) - u(A)\| \leq -\frac{\|\Delta A|}{\sigma_1(\Delta A)} \log\left(1 - \frac{\sigma_1(\Delta A)}{\sigma_n(A)}\right).$$

THEOREM 3.3. *Let $A, \Delta A \in M_n(C)$ be given. Assume that $A+t\Delta A$ is nonsingular for all $t \in [0,1]$. Then for any unitarily invariant norm $|||\cdot|||$*

$$(3.1) \qquad |||U(A+\Delta A) - U(A)||| \leq \max\{[\sigma_n(A+t\Delta A)]^{-1} : 0 \leq t \leq 1\}\ |||\Delta A|||.$$

COROLLARY 3.4. *Let $A \in M_n(C)$ be nonsingular. Then the condition of the function $U(\cdot)$ at A with respect to any unitarily invariant norm is*

$$(3.2) \qquad \kappa(A) = \sigma_n^{-1}(A).$$

Proof. Use Theorem 3.2 or 3.3 to show $\kappa(A) \leq \sigma_n^{-1}(A)$. For the reverse inequality use the method on the proof of Corollary 2.5 with

$$B = U(0_{n-1} \oplus [1])V. \qquad \square$$

THEOREM 3.5. *Let $A \in M_n(C)$ be nonsingular and let $\Delta A \in M_n(C)$ be such that $\sigma_1(\Delta A) < \sigma_n(A)$. Then we have the majorization*

(3.3)

$$\sum_{i=1}^{k} \sigma_i(U(A + \Delta A) - U(A)) \leq -\sum_{i=1}^{k} \log\left(1 - \frac{\sigma_i(\Delta A)}{\sigma_{n+1-i}(A)}\right), \qquad k = 1, \ldots, n$$

and the bound

(3.4)

$$|||U(A + \Delta A) - U(A)||| \leq g\left(\log\left(1 - \frac{\sigma_1(\Delta A)}{\sigma_n(A)}\right), \ldots, \log\left(1 - \frac{\sigma_n(\Delta A)}{\sigma_1(A)}\right)\right),$$

where g is the symmetric gauge function corresponding to $||| \cdot |||$.

4. Counterexample to a best approximation conjecture. In this section we will consider the unitary Procrustes problem:

$$\min\{\|A - YB\| : Y \in M_n(C), Y^*Y = I\}, \tag{4.1}$$

where $A, B \in M_n(C)$ and a unitarily invariant norm $\| \cdot \|$ are given. It is easy to show that if $\| \cdot \|$ is the Frobenius norm then Y attains the minimum in (4.1) if and only if Y^*BA^* is positive semidefinite (see, e.g., [5, Example 7.4.8] for the details). If A and B are nonsingular then this says that $U(AB^*)$ is the unique solution to the unitary Procrustes problem (4.1) for the Frobenius norm. Marshall and Olkin asked whether this is true for all unitarily invariant norms [9, pp. 269–270]. We show that it is not true, and solve problem (4.1) for the trace norm ($||| \cdot |||_n$) in a special case.

Our analysis is based on the following lemma [5, Thm. 7.4.9].

LEMMA 4.1. *Let $A \in M_n(C)$ be given. Then*

$$\text{Re tr } YA \leq \text{tr } |A| = |||A|||_n \quad \textit{for all unitary } Y \in M_n(C)$$

with equality if and only if UA is positive semidefinite.

LEMMA 4.2. *Let $A, B \in M_n(C)$ be given and assume that $|A| - |B|$ is positive semidefinite. Let $V, W \in M_n(C)$ be unitary and such that $V^*A = |A|$ and $W^*B = |B|$. Then*

(4.2)

$$|||A - YB|||_n \geq |||A - VW^*B|||_n = \text{tr}(|A| - |B|) \quad \textit{for all unitary } Y \in M_n(C).$$

Equality holds if and only if $Y = V_1W_1^*$ where V_1 and W_1 are unitary matrices such that $V_1^*A = |A|$ and $W_1^*B = |B|$.

Proof. Since $|A| - |B|$ is positive semidefinite we have

$$|||A - VW^*B|||_n = |||V^*A - W^*B|||_n = |||\,|A| - |B|\,|||_n = \text{tr }(|A| - |B|).$$

Now given any unitary $Y \in M_n(C)$ and any unitary $V_1 \in M_n(C)$ such that $V_1^*A = |A|$, let $W_1 = V_1^*U$. Then using the triangle inequality and Lemma 4.1 we have

$$\begin{aligned} |||A - YB|||_n &= |||\, |A| - V_1^*YB \,|||_n \\ &= |||\, |A| - W_1B \,|||_n \\ &\geq |||\, |A| \,|||_n - |||W_1B|||_n \\ &\geq \operatorname{tr} |A| - \operatorname{tr} |B| \qquad (4.3) \\ &= \operatorname{tr} (|A| - |B|), \end{aligned}$$

which establishes (4.2).

All that remains is to show that if equality holds in (4.2) then Y must be of the required form. If Y is such that equality holds in (4.2) then we must have equality in (4.3). By Lemma 4.1 this requires that W_1B be positive semidefinite as desired. □

Now we will show that $U(BA^*)$ is not a solution to (4.1) for the trace norm if

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

It is clear that $A, B \in M_n(C)$ are positive semidefinite and that $A - B$ is also positive semidefinite. Since A and B do not commute, BA^* is not positive semidefinite and so $U(BA^*) \neq I$. But from Lemma 4.2 we see that I is the unique solution to (4.1) for the trace norm.

5. Bounds on $\|\, |A| - |B| \,\|$. Considerably less is known about bounds of the form

$$\|\, |A| - |B| \,\| \leq c\|A - B\|,$$

where $\|\cdot\|$ is a given unitarily invariant norm and c is a constant, possibly depending on $\|\cdot\|$. In fact, little is known except for the Frobenius norm (in which case the problem has been solved) and the operator norm (in which case there are some partial results). Davies [3] has obtained bounds for the Schatten p-norms.

THEOREM 5.1. *Let $n \geq 2$ and let $A, B \in M_n(C)$. Then*

$$\|\, |A| - |B| \,\|_F \leq \sqrt{2}\|A - B\|_F. \qquad (5.1)$$

Furthermore, $\sqrt{2}$ is the best possible constant.

Proof. The inequality (5.1) is stated in [2] with the additional restriction

$$\sigma_n(A) + \sigma_{n-1}(A) > \sqrt{2}\|A - B\|_F.$$

A slight modification of the proof in [2] shows that this restriction is not necessary. Alternatively, see [3, Lemma 2].

To see that $\sqrt{2}$ is the best possible constant take

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & \epsilon \\ 0 & 0 \end{pmatrix},$$

and let $\epsilon \downarrow 0$. □

From Theorem 5.1 it follows that we have a bound for the operator norm:

$$|||\, |A| - |B| \,|||_1 \leq c_n|||A - B|||_1 \quad \text{for all } A, B \in M_n(C), \quad c_n = \sqrt{2n}. \qquad (5.2)$$

In [7] Kato shows that there is no constant c independent of n such that

$$|||\,|A| - |B|\,|||_1 \leq c|||A - B|||_1 \quad \text{for all } A, B \in M_n(C).$$

It appears that the best value of c_n for each n is unknown. It is known that

$$\gamma_n \leq c_n \leq 2\gamma_n + 1,$$

where $\gamma_n \approx (2/\pi) \log n$ is a known constant [10, Cor. 4.3].

REFERENCES

[1] T. Ando, R. A. Horn, and C. R. Johnson, *The singular values of a Hadamard product: A basic inequality*, Linear and Multilinear Algebra, 21 (1987), pp. 345–365.

[2] A. Barrlund, *Perturbation bounds on the polar decomposition*, BIT, 30 (1989), pp. 101–113.

[3] E. B. Davies, *Lipschitz continuity of functions of operators in the Schatten classes*, J. London Math. Soc., 37 (1988), pp. 148–157.

[4] N. J. Higham, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.

[5] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[6] ———, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[7] T. Kato, *Continuity of the map $S \rightarrow |S|$ for linear operators*, Proc. Japan Acad., 49 (1973), pp. 157–160.

[8] C. Kenney and A. J. Laub, *Polar decomposition and matrix sign function condition estimates*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 488–504.

[9] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, London, 1979.

[10] R. Mathias, *The Hadamard operator norm of a circulant and applications*, SIAM J. Matrix Anal. Appl., 14 (1993), to appear.

[11] ———, *The singular values of the Hadamard product of a positive semidefinite and a skew-symmetric matrix*, Linear and Multilinear Algebra, 31 (1992), pp. 57–70.

[12] R. Schrieber and B. Parlett, *Block reflectors: Theory and computation*, SIAM J. Numer. Anal., 25 (1988), pp. 189–205.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 2, pp. 598–618, April 1993

ERROR ANALYSIS OF UPDATE METHODS FOR THE SYMMETRIC EIGENVALUE PROBLEM*

JESSE L. BARLOW†

Abstract. Cuppen's divide-and-conquer method for solving the symmetric tridiagonal eigenvalue problem has been shown to be very efficient on shared memory multiprocessor architectures.

In this paper, some error analysis issues concerning this method are resolved. The method is shown to be stable and a slightly different stopping criterion for finding the zeroes of the spectral function is suggested.

These error analysis results extend to general update methods for the symmetric eigenvalue problem. That is, good backward error bounds are obtained for methods to find the eigenvalues and eigenvectors of $A + \rho ww^T$, given those of A. These results can also be used to analyze a new fast method for finding the eigenvalues of banded, symmetric Toeplitz matrices.

Key words. divide-and-conquer, spectral function, eigenvalue update, Toeplitz matrices

AMS(MOS) subject classifications. 65F15, 65G05

1. Introduction.

In this paper, we provide an error analysis of Cuppen's [4] divide-and-conquer algorithm for solving the symmetric tridiagonal eigenvalue problem. The method is highly suitable for shared memory multiprocessor computers [8], [15]. We make some implementation suggestions that are slightly different from those proposed by Dongarra and Sorensen [8]. Jessup and Sorensen [16], [15] developed a variant of the procedure for the bidiagonal singular value decomposition. Although we do not explicitly analyze this method, the ideas in this paper could be extended to the Jessup–Sorensen procedure. These algorithms are being considered for implementation in LAPACK [5].

In fact, this analysis really concerns more general updating strategies arising out of the eigenvalue update procedure of Bunch, Nielsen, and Sorensen [3]. Robust implementations of this procedure have been proposed by Sorensen and Tang [23] and by Kahan [18]. For the most part, we will discuss the Sorensen–Tang [23] implementation. This analysis is also used to prove the stability of the procedure of Handy and Barlow [13] for symmetric, banded Toeplitz matrices.

The paper is organized as follows. In §2, we summarize Dongarra and Sorensen's version of the divide-and-conquer algorithm of Cuppen [4] and extend some of the results there. In §3, we discuss the problem of finding the zeroes of the spectral function, and give the properties of a more general stopping criterion. Section 4 presents a backward error analysis of the algorithm and of general update procedures. In §5, we discuss the implications of our results for the Toeplitz eigenvalue procedure of Handy and Barlow [13], which we demonstrate. Tests by Jessup and Ipsen [14], [17], [15] indicate that Cuppen's method obtains eigenvectors that are more closely orthogonal than those obtained by inverse iteration and obtain eigenvalues that are accurate in the absolute sense given in [24], [20], and [10]. Whether or not the method

* Received by the editors September 9, 1991; accepted for publication November 5, 1991.

† Department of Computer Science, The Pennsylvania State University, University Park, Pennsylvania 16802-6103 (barlow@cs.psu.edu). This research was supported in part by Air Force Office of Scientific Research grant AFOSR-88-0161 and by National Science Foundation grant CCR-9000526. This research was done in part while the author was visiting Oak Ridge National Laboratory under the support of the Applied Mathematical Sciences Research Program of the Office of Energy Research, U.S. Department of Energy, under contract DE-AC05-840R21400 with Martin Marietta Energy Systems, Inc.

obtains good eigenvalues in the strong relative error sense defined by Barlow and Demmel [1] is an open question.

2. An outline of Cuppen's method for the symmetric tridiagonal eigenproblem. Consider the problem of finding the eigenvalues of the symmetric tridiagonal matrix T given by

$$T = \begin{pmatrix} a_1 & b_1 & \cdot & 0 & 0 \\ b_1 & a_2 & b_2 & \cdot & \cdot \\ \cdot & b_2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & b_{n-1} \\ \cdot & \cdot & \cdot & b_{n-1} & a_n \end{pmatrix}.$$

We partition the problem by rank-one tearing. To do that, we consider the partitioning

$$T = \begin{pmatrix} T_1 & b_k e_k e_1^T \\ b_k e_1 e_k^T & T_2 \end{pmatrix},$$

where $1 \leq k \leq n$ and e_j represents the jth unit vector of appropriate dimension. We then write T as

$$T = \begin{pmatrix} \hat{T}_1 & 0 \\ 0 & \hat{T}_2 \end{pmatrix} + \rho w w^T,$$

$$\rho = \hat{\theta} b_k, \quad w = (1+\theta^{-2})^{-\frac{1}{2}} \begin{pmatrix} e_k \\ \theta^{-1} e_1 \end{pmatrix}, \quad \theta \neq 0, \quad \hat{\theta} = \theta\sqrt{1+\theta^{-2}}, \tag{1}$$

where $\hat{T}_1$ and $\hat{T}_2$ are modifications of T_1 and T_2 such that

$$\hat{T}_1 = \begin{pmatrix} a_1 & b_1 & 0 & \cdot & 0 & 0 \\ b_1 & a_2 & b_2 & \cdot & \cdot & \cdot \\ \cdot & b_2 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & b_{k-2} & a_{k-1} & b_{k-1} \\ 0 & \cdot & \cdot & \cdot & b_{k-1} & \tilde{a}_k \end{pmatrix}, \tag{2}$$

$$\hat{T}_2 = \begin{pmatrix} \tilde{a}_{k+1} & b_{k+1} & 0 & \cdot & 0 & 0 \\ b_{k+1} & a_{k+2} & b_{k+2} & \cdot & \cdot & \cdot \\ \cdot & b_{k+2} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & b_{n-2} & a_{n-1} & b_{n-1} \\ 0 & \cdot & \cdot & \cdot & b_{n-1} & a_n \end{pmatrix}. \tag{3}$$

Here

$$\tilde{a}_k = a_k - \theta b_k, \qquad \tilde{a}_{k+1} = a_{k+1} - \theta^{-1} b_k. \tag{4}$$

The parameter θ is chosen to avoid potential numerical difficulties associated with cancellation. If $\text{sign}(a_k) = \text{sign}(a_{k+1})$, we choose $\theta = \pm 1$ so that $\text{sign}(-\theta b_k) = \text{sign}(a_k)$. If $\text{sign}(a_k) = -\text{sign}(a_{k+1})$, Dongarra and Sorensen [8] recommend that "the sign of θ" be "chosen so that θb_k has the same sign as one of the elements and the magnitude of θ" be "chosen to avoid severe loss of significant digits when $\theta^{-1} b_k$ is subtracted from the other." Actually, for this analysis, we can always just choose $\theta = \text{sign}(a_k)$. The value $\theta = \pm 1$ will be assumed for the rest of this paper.

The tearing operation (3)–(4) separates the problem of finding the eigenvalues of T into that of solving the two eigensystems

$$\hat{T}_1 = Q_1 D_1 Q_1^T, \qquad \hat{T}_2 = Q_2 D_2 Q_2^T,$$

where Q_1 and Q_2 are orthogonal matrices and D_1 and D_2 are diagonal matrices. These two eigenproblems can be assigned to separate sets of processors. The eigensystems of $\hat{T}_1$ and $\hat{T}_2$ can each be "torn" into smaller eigensystems in the same manner. This process can be performed recursively down to a 1×1 or 2×2 eigensystem, hence the divide-and-conquer nature of the algorithm. The eigenvalues of T can be obtained from

$$\begin{aligned} T &= \begin{pmatrix} Q_1 D_1 Q_1^T & 0 \\ 0 & Q_2 D_2 Q_2^T \end{pmatrix} + \theta b_k \begin{pmatrix} e_k \\ \theta^{-1} e_1 \end{pmatrix} (e_k^T, \theta^{-1} e_1^T) \\ &= \text{diag}(Q_1, Q_2) \left[\text{diag}(D_1, D_2) + \theta b_k \begin{pmatrix} q_1 \\ \theta^{-1} q_2 \end{pmatrix} (q_1^T, \theta^{-1} q_2^T) \right] \text{diag}(Q_1^T, Q_2^T), \end{aligned}$$

where $q_1 = Q_1^T e_k$ and $q_2 = Q_2^T e_1$. We can then find the eigenvalues of T by solving the eigenproblem

$$D + \rho z z^T = \hat{Q} \hat{D} \hat{Q}^T,$$

where

$$D = \text{diag}(D_1, D_2), \quad \rho = \hat{\theta} b_k, \quad z = (1 + \theta^{-2})^{-\frac{1}{2}} \begin{pmatrix} q_1 \\ \theta^{-1} q_2 \end{pmatrix}. \tag{5}$$

We note that we have normalized z in (5) and w in (1) so that $\|z\|_2 = \|w\|_2 = 1$. This will be assumed throughout this paper. Here $\hat{D}$ is a diagonal matrix of the eigenvalues and $\hat{Q}$ is the orthogonal matrix of eigenvectors. Let $D = \text{diag}(\delta_1, \delta_2, \ldots, \delta_n)$ where $\delta_1 > \delta_2 > \cdots > \delta_n$ and assume that no component ζ_i of z is zero. If (q, λ) is an eigenpair, then

$$(D + \rho z z^T) q = \lambda q$$

and thus

$$(D - \lambda I) q = -\rho (z^T q) z.$$

We first show that $(D - \lambda I)^{-1}$ exists. If $\lambda = \delta_i$ for some i, then our assumption that $\zeta_i \neq 0$ implies that $z^T q = 0$. This, together with the assumption of distinct eigenvalues, implies that q is the eigenvector e_i of D; then $z^T e_i = 0$ contradicts the original assumption. Thus we have established the existence of $(D - \lambda I)^{-1}$. Some algebraic manipulations imply that an eigenvalue λ is a root of the equation

$$f(\lambda) = 0, \tag{6}$$

where

$$f(\lambda) = 1 + \rho z^T (D - \lambda I)^{-1} z = 1 + \rho \sum_{i=1}^{n} \frac{\zeta_j^2}{\delta_j - \lambda}. \tag{7}$$

The function f in (7) is often referred to as the spectral function. Golub [9] first discussed the problem of finding the roots of (6)–(7). Without loss of generality, we can assume that $\rho > 0$. If $\rho < 0$, we consider $-D - \rho z z^T$.

Bunch, Nielsen, and Sorensen [3] show that the eigenvectors, the columns of $\hat{Q}$, are given by

$$\hat{q}_i = \gamma_i (D - \lambda_i I)^{-1} z, \tag{8}$$

where γ_i is chosen to make $\|\hat{q}_i\|_2 = 1$ and λ_i is the ith eigenvalue of $D + \rho z z^T$.

In the case where δ_i and δ_{i+1} are close, Dongarra and Sorensen [8] advocate a deflation strategy. Essentially, if $|\delta_{i+1} - \delta_i|$ is small, we consider the 2×2 submatrix of $D + \rho z z^T$ given by

$$\begin{pmatrix} \delta_i & 0 \\ 0 & \delta_{i+1} \end{pmatrix} + \rho \begin{pmatrix} \zeta_i \\ \zeta_{i+1} \end{pmatrix} (\zeta_i, \zeta_{i+1}).$$

We can construct a Givens rotation to introduce a zero in one of the two corresponding components of z. It is given by

$$\begin{aligned} &\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \left[\begin{pmatrix} \delta_i & 0 \\ 0 & \delta_{i+1} \end{pmatrix} + \rho \begin{pmatrix} \zeta_i \\ \zeta_{i+1} \end{pmatrix} (\zeta_i, \zeta_{i+1}) \right] \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \\ &\quad = \begin{pmatrix} \hat{\delta}_i & 0 \\ 0 & \hat{\delta}_{i+1} \end{pmatrix} + \rho \begin{pmatrix} \tau \\ 0 \end{pmatrix} (\tau, 0) + (\delta_{i+1} - \delta_i) cs \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \end{aligned}$$

where $c^2 + s^2 = 1$, $\hat{\delta}_i = \delta_i c^2 + \delta_{i+1} s^2$, $\hat{\delta}_{i+1} = \delta_i s^2 + \delta_{i+1} c^2$, and $\tau = \zeta_i^2 + \zeta_{i+1}^2$. If we put G_i equal to the $n \times n$ Givens transformation constructed from the identity by replacing the appropriate diagonal block with the 2×2 rotation just constructed, we have

$$G_i (D + \rho z z^T) G_i^T = \bar{D} + \bar{z}\bar{z}^T + E_i,$$

where $e_i^T \bar{z} = \tau$ and $e_{i+1}^T \hat{z} = 0$, $e_i^T \bar{D} e_i = \delta_i$, $e_{i+1}^T \bar{D} e_{i+1} = \delta_{i+1}$, and

$$\|E_i\|_2 = |(\delta_{i+1} - \delta_i) cs|.$$

Deflation is done only when $|\delta_i - \delta_{i+1}||cs| \le \epsilon_2$ is satisfied where ϵ_2 is some tolerance. Deflation can speed up this algorithm considerably.

Dongarra and Sorensen [8] show that the error bound caused by deflation will be no more than that caused by a perturbation δT_D where $\|\delta T_D\|_2 \le u p_D(n) \|T\|_2$ and $p_D(n)$ is some modestly growing function of n. It is easy to show that the reductions from rank-one tearing, that is, the arithmetic in obtaining $\hat{T}_1$ and $\hat{T}_2$ from T, contribute a backward error δT_R where

$$\|\delta T_R\|_2 \le \bar{c} u \|T\|_2 + O(u^2), \qquad \bar{c} = O(1).$$

Thus, in the absolute error sense, neither the reduction nor the deflation will seriously affect the quality of the eigenvalues.

The remaining problem is to show that computing the eigenvalues from (6)–(7) produces a backward error of no more than $\|\delta T_S\|_2$ where

$$\|\delta T_S\|_2 \le u p_S(n) \|T\|_2 + O(u^2).$$

Here $p_S(n)$ is a modestly growing function of n.

In the next section, we discuss how to find the zeroes of $f(\lambda)$ in (7).

3. Finding the zeroes of the spectral function. Bunch, Nielsen, and Sorensen [3] developed a method inspired by the work of Moré [19] and Reinsch [21], [22] to find the zeroes of f in (7). It relies upon rational approximation to construct an iterative method to solve (6)–(7). To find a zero of f, we write the function as

$$f(\lambda) = 1 + \phi(\lambda) + \psi(\lambda),$$

where

$$\psi(\lambda) = \sum_{j=1}^{i} \frac{\zeta_j^2}{(\delta_j - \lambda)}$$

and

$$\phi(\lambda) = \sum_{j=i+1}^{n} \frac{\zeta_j^2}{(\delta_j - \lambda)}.$$

From a standard result on the symmetric eigenvalue problem [10, Chap. 8], the root λ_i lies in the open interval (δ_i, δ_{i+1}). For λ in that interval, all terms of $\psi(\lambda)$ are negative and all terms of $\phi(\lambda)$ are positive. An iterative method is derived for solving the equation

$$-\psi(\lambda) = 1 + \phi(\lambda) \tag{9}$$

by starting with an initial guess $\lambda_i^{(0)} \in (\delta_i, \delta_{i+1})$, and constructing simple rational interpolants of the form

$$\frac{d_0}{d_1 - \lambda}, \qquad d_2 + \frac{d_3}{\delta - \lambda}, \tag{10}$$

where δ is fixed at the current value of δ_{i+1} and the parameters $d_i, i = 0,1,2,3$ are defined by the interpolation conditions

$$\frac{-d_0}{d_1 - \lambda_i^{(0)}} = \psi(\lambda_i^{(0)}), \qquad d_2 + \frac{d_3}{\delta - \lambda_i^{(0)}} = \phi(\lambda_i^{(0)}); \tag{11}$$

$$\frac{d_0}{(d_1 - \lambda_i^{(0)})^2} = \psi'(\lambda_i^{(0)}), \qquad \frac{d_3}{(\delta - \lambda_i^{(0)})^2} = \phi'(\lambda_i^{(0)}). \tag{12}$$

The new approximate $\lambda_i^{(1)}$ is the root of the equation

$$\frac{-d_0}{d_1 - \lambda} = 1 + d_2 + \frac{d_3}{\delta - \lambda}. \tag{13}$$

We can construct an initial guess that lies in the open interval (δ_i, λ_i). A sequence of iterates $\{\lambda_i^{(k)}\}$ can be constructed where $\lambda_i^{(k+1)}$ is derived from $\lambda_i^{(k)}$ as $\lambda_i^{(1)}$ was from $\lambda_i^{(0)}$. Bunch, Nielsen, and Sorensen [3] show that this iteration is quadratically convergent and the iterates satisfy $\lambda_i^{(k)} < \lambda_i^{(k+1)} < \lambda_i$, thus there is no need for safeguarding. The stopping criteria used are

$$|f(\lambda_i^{(k+1)})| \leq \max\{|\delta_1|, |\delta_n|\}\epsilon_1, \tag{14}$$

$$|\lambda_i^{(k+1)} - \lambda_i^{(k)}| \le \min\{|\delta_i - \lambda_i^{(k+1)}|, |\delta_{i+1} - \lambda_i^{(k+1)}|\}\epsilon_1, \tag{15}$$

where $\lambda_i^{(k+1)}$ is the current iterate. The second condition is designed to preserve the orthogonality of the eigenvectors. The stopping criterion (14) does not seem necessary from the analysis in this paper.

We instead propose that the single stopping criterion

$$\frac{|f(\lambda_i^{(k)})|}{|f'(\lambda_i^{(k)})|} \le \min\{|\delta_i - \lambda_i^{(k)}|, |\delta_{i+1} - \lambda_i^{(k)}|\}\epsilon_1 \tag{16}$$

be used.

Actually, in exact arithmetic (15) and (16) are nearly equivalent, as shown by the following theorem.

THEOREM 3.1. *Let the stopping criterion*

$$|\lambda_i^{(k+1)} - \lambda_i^{(k)}| \le \min\{|\delta_i - \lambda_i^{(k)}|, |\delta_{i+1} - \lambda_i^{(k)}|\}\epsilon_1$$

be satisfied by the iterates of the Bunch, Nielsen, and Sorensen procedure. (Note that the right side uses $\lambda_i^{(k)}$ and not $\lambda_i^{(k+1)}$.) Then if $\epsilon_1 \le 0.2290$, in the absence of rounding error,

$$\frac{2}{3}\frac{|f(\lambda_i^{(k)})|}{|f'(\lambda_i^{(k)})|} \le |\lambda_i^{(k+1)} - \lambda_i^{(k)}| \le 2\frac{|f(\lambda_i^{(k)})|}{|f'(\lambda_i^{(k)})|}.$$

Before proving this theorem, we need two simple lemmas.

LEMMA 3.2. *For the current iterate $\lambda_i^{(k)}$, the coefficient d_1 in (10)–(11) satisfies*

$$|d_1 - \lambda_i^{(k)}| \ge |\delta_i - \lambda_i^{(k)}|.$$

Proof. Let $\mu = \lambda_i^{(k)}$. From the interpolation property,

$$\frac{-d_0}{d_1 - \mu} = \psi(\mu), \qquad \frac{-d_0}{(d_1 - \mu)^2} = \psi'(\mu).$$

If we eliminate d_0 from these expressions, then we get the quadratic equation in $d_1 - \lambda$ given by

$$(d_1 - \mu)\psi(\mu) - (d_1 - \mu)^2\psi'(\mu) = 0.$$

This has two roots, $d_1 - \mu = 0$, which violates the interpolation condition by creating a pole at μ, and

$$d_1 - \mu = \frac{\psi(\mu)}{\psi'(\mu)}. \tag{17}$$

Let $\bar{D} = \operatorname{diag}(\delta_1, \dots, d_i)$ and let $\bar{z} = (z_1, z_2, \dots, z_i)^T$. Then

$$\psi(\mu) = \rho\bar{z}^T(\bar{D} - \mu I)^{-1}\bar{z}, \qquad \psi'(\mu) = \rho\bar{z}^T(\bar{D} - \mu I)^{-2}\bar{z}.$$

We have

$$|d_1 - \mu| = \frac{|\rho\bar{z}^T(\bar{D} - \mu I)^{-1}\bar{z}|}{|\rho\bar{z}^T(\bar{D} - \mu I)^{-2}\bar{z}|} = \frac{|y^T(\bar{D} - \mu I)y|}{y^T y} \ge |\delta_i - \mu|,$$

where

$$y = (D - \mu I)^{-1} z. \qquad \square$$

We need a second lemma to establish the signs of d_0 and d_3.

LEMMA 3.3. *The coefficients d_0 in* (10) *and d_3 in* (12) *satisfy $d_0 \leq 0$ and $d_3 \geq 0$.*

Proof. From the interpolation condition,

$$\frac{-d_0}{d_1 - \lambda} = \psi(\lambda) \geq 0.$$

From the form of $\psi'(\lambda)$ and (17),

$$d_1 - \lambda = \frac{\psi(\lambda)}{\psi'(\lambda)} \geq 0.$$

Thus $d_0 \geq 0$. Likewise we have from the form of $\phi'(\lambda)$,

$$\phi'(\lambda) \geq 0.$$

Thus

$$d_3 = (\delta_{i+1} - \lambda)^2 \phi'(\lambda) \geq 0. \qquad \square$$

Now we can prove Theorem 3.1.

Proof of Theorem 3.1. Let $\mu = \lambda_i^{(k)}$ and $\hat{\mu} = \lambda_i^{(k+1)}$. The rational interpolation function is

$$\tau(\lambda) = 1 + d_2 + \frac{d_0}{d_1 - \lambda} + \frac{d_3}{\delta_{i+1} - \lambda}.$$

From the interpolation property we have

$$\tau(\mu) = f(\mu), \qquad \tau'(\mu) = f'(\mu),$$

and from the fact that $\hat{\mu}$ is a root we have $\tau(\hat{\mu}) = 0$. Using a Taylor expansion of $\tau(\hat{\mu})$ yields

$$0 = \tau(\hat{\mu}) = f(\mu) + (\hat{\mu} - \mu) f'(\mu) + \frac{(\hat{\mu} - \mu)^2}{2} \tau''(\omega),$$

where ω is between $\hat{\mu}$ and μ. Therefore,

$$\frac{f(\mu)}{f'(\mu)} = (\hat{\mu} - \mu) + \frac{(\hat{\mu} - \mu)^2}{2} \frac{\tau''(\omega)}{f'(\mu)}. \tag{18}$$

Noting that $d_0 \leq 0$ and $d_3 \geq 0$ from Lemma 3.3, we have that

$$\begin{aligned} \tau'(\lambda) &= \frac{-d_0}{(d_1 - \lambda)^2} + \frac{d_3}{(\delta_{i+1} - \lambda)^2} = d^T (F - \lambda I)^{-2} d, \\ \tau''(\lambda) &= \frac{-2d_0}{(d_1 - \lambda)^3} + \frac{2d_0}{(\delta_{i+1} - \lambda)^3} = 2d^T (F - \lambda I)^{-3} d, \\ d &= (\sqrt{-d_0}, \sqrt{d_3})^T, \qquad F = \operatorname{diag}(d_1, \delta_{i+1}). \end{aligned}$$

Thus

$$\frac{\tau''(\omega)}{f'(\mu)} = 2\frac{d^T(F-\omega I)^{-3}d}{d^T(F-\mu I)^{-2}d} = 2\frac{y^T(F-\omega I)^{-3}(F-\mu I)^2 y}{y^T y},$$
$$y = (F-\mu I)^{-1}d.$$

Therefore, we have the bound

$$\frac{|\tau''(\omega)|}{|f'(\mu)|} \le 2\|(F-\omega I)^{-3}(F-\mu I)^2\|_2$$
$$= 2\max\left\{\frac{(d_1-\mu)^2}{|d_1-\omega|^3}, \frac{(\delta_{i+1}-\mu)^2}{|\delta_{i+1}-\omega|^3}\right\}.$$

The second term in (18) satisfies

$$\frac{|\tau''(\omega)|}{2|f'(\mu)|}(\hat{\mu}-\mu)^2 \le (\hat{\mu}-\mu)^2 \max\left\{\frac{(d_1-\mu)^2}{|d_1-\omega|^3}, \frac{(\delta_{i+1}-\mu)^2}{|\delta_{i+1}-\omega|^3}\right\}$$
$$\le |\hat{\mu}-\mu| \max\left\{\frac{(d_1-\mu)^2}{|d_1-\omega|^3}, \frac{(\delta_{i+1}-\mu)^2}{|\delta_{i+1}-\omega|^3}\right\} \min\{|\delta_i-\mu|, |\delta_{i+1}-\mu|\}\,\epsilon_1$$
$$\le |\hat{\mu}-\mu| \max\left\{\frac{|d_1-\mu|^3}{|d_1-\omega|^3}, \frac{|\delta_{i+1}-\mu|^3}{|\delta_{i+1}-\omega|^3}\right\}\epsilon_1.$$

Since $|\omega-\mu| \le |\hat{\mu}-\mu| \le \min\{|\delta_i-\mu|, |\delta_{i+1}-\mu|\}\epsilon_1$, we have

$$\frac{|\tau''(\omega)|}{|f'(\mu)|}|\hat{\mu}-\mu|^2 \le |\hat{\mu}-\mu|\left(\frac{1}{1-\epsilon_1}\right)^3\epsilon_1.$$

Let $s = (1/(1-\epsilon_1))^3\epsilon_1$. Substituting back into (18) yields

$$\frac{1}{1+s}\frac{|f(\mu)|}{|f'(\mu)|} \le |\hat{\mu}-\mu| \le \frac{1}{1-s}\frac{|f(\mu)|}{|f'(\mu)|}.$$

If $s < \frac{1}{2}$ then the theorem is satisfied. We may verify that if $\epsilon_1 < .2290$ then $s < \frac{1}{2}$. □

We should note that if $\lambda_i^{(k+1)}$ and $\lambda_i^{(k)}$ satisfy (15) then

$$|\lambda_i^{(k+1)} - \lambda_i^{(k)}| \le \frac{1}{1-\epsilon_1}\min\{|\delta_{i+1}-\lambda_i^{(k)}|, |\delta_{i+1}-\lambda_i^{(k)}|\}\epsilon_1.$$

The condition (16) is used for the rest of this paper for two reasons.

1. It is relatively independent of the method used to find the zeroes of the spectral function. The approaches of both Sorensen and Tang [23] and Kahan [18] are based upon interpolating the function and its derivative, therefore, (16) can be computed with little extra effort.

2. The results given below work nicely with (16). In the absence of rounding error, they would work similarly for (15) with the Bunch, Nielsen, and Sorensen [3] iteration. We see no reason why (15) cannot be used even in the presence of rounding error, although we know of no results in this direction.

It is important to show that the stopping criterion (16) is effectively satisfied. The following theorem generalizes a discussion in Sorensen and Tang [23]. This discussion was inspired by Kahan [18]. The difference here is that we do not require $f(\lambda)$ to be evaluated at the root.

THEOREM 3.4. *Let $\mu = \lambda_i^{(k)}$. Let $f(\mu)$ be computed on a floating point computer with machine unit u_2, let $f'(\mu)$ be computed with machine unit $u \geq u_2$, and let $f(\mu)/f'(\mu)$ be computed with machine unit u. Assume*

$$\textit{computed}\left(\frac{|f(\mu)|}{|f'(\mu)|}\right) \leq \|(D-\mu I)^{-1}\|_2^{-1}\epsilon_1. \tag{19}$$

Then

$$\frac{|f(\mu)|}{|f'(\mu)|} \leq \|(D-\mu I)^{-1}z\|_2^{-1}(2n+8)u_2 + \|(D-\mu I)^{-1}\|_2^{-1}\hat{\epsilon_1} + O(u_2\epsilon_1), \tag{20}$$

where

$$|\hat{\epsilon_1}| \leq \epsilon_1(1+(n+5)u) + O(\epsilon_1 u^2).$$

Proof. Let $fl(\cdot)$ describe computations in machine unit u and let $fl_2(\cdot)$ describe computations in machine unit u_2.

First, we note that

$$f(\mu) = 1 + \rho z^T(D-\mu I)^{-1}z = 1 + \rho\sum_{i=1}^{n}\frac{\zeta_j^2}{\delta_j - \mu},$$
$$f'(\mu) = \rho z^T(D-\mu I)^{-2}z = \rho\sum_{i=1}^{n}\frac{\zeta_j^2}{(\delta_j-\mu)^2}.$$

From standard error analysis results for sums and inner products [24], we have

$$\begin{gathered} fl_2(f(\mu)) = f(\mu) + (1 + |z|^T|(D-\mu I)^{-1}z|)\sigma_0, \\ fl(f'(\mu)) = f'(\mu)(1+\sigma_1) = \rho\|(D-\mu I)^{-1}z\|_2^2(1+\sigma_1), \\ |\sigma_0| \leq (n+4)u_2 + O(u_2^2), \qquad |\sigma_1| \leq (n+4)u + O(u^2). \end{gathered} \tag{21}$$

Here $|z|$ denotes componentwise absolute value. To simplify (21), we use the fact that $|z|^T|(D-\mu I)^{-1}z| \leq \|z\|_2\|(D-\mu I)^{-1}z\|_2$, and the fact that $\|z\|_2 = 1$. We then have

$$\text{computed}\left(\frac{|f(\mu)|}{|f'(\mu)|}\right) \leq \frac{|f(\mu)|}{|f'(\mu)|}(1+\sigma_2) - \frac{1+\rho\|(D-\mu I)^{-1}z\|_2}{\rho\|(D-\mu I)^{-1}z\|_2^2}\sigma,$$
$$|\sigma_2| \leq (n+5)u + O(u^2), \qquad |\sigma| \leq |\sigma_0| \leq (n+5)u_2 + O(u_2^2).$$

From the assumption (19), we have that

$$\frac{|f(\mu)|}{|f'(\mu)|} \leq \frac{1+\rho\|(D-\mu I)^{-1}z\|_2}{\rho\|(D-\mu I)^{-1}z\|_2^2}\sigma + \|(D-\mu I)^{-1}\|_2^{-1}\hat{\epsilon_1}, \tag{22}$$

where

$$\hat{\epsilon_1} = \epsilon_1(1+\sigma_2)^{-1} \leq \epsilon_1(1+(n+5)u) + O(u^2\epsilon_1).$$

Using the expression for $f(\mu)$ and $f'(\mu)$ and (22), we have

$$\begin{aligned} 1 &\leq -\rho z^T(D-\mu I)^{-1}z + (1+\rho\|(D-\mu I)^{-1}z\|_2)\sigma \\ &\quad + \rho\|(D-\mu I)^{-1}\|_2^{-1}\|(D-\mu I)^{-1}z\|_2^2\hat{\epsilon_1} \\ &\leq \|(D-\mu I)^{-1}z\|_2 + \sigma + \rho\|(D-\mu I)^{-1}z\|_2(\sigma+\hat{\epsilon_1}), \end{aligned}$$

from $\|(D-\mu I)^{-1}z\|_2 \leq \|(D-\mu I)^{-1}\|_2\|z\|_2$ and the fact that $\|z\|_2 = 1$. Therefore,

$$1 \leq \rho\|(D-\mu I)^{-1}z\|_2(1+\sigma+\hat{\epsilon_1})/(1-\sigma). \tag{23}$$

If we put (23) into (22), we have

$$\begin{aligned}\frac{|f(\mu)|}{|f'(\mu)|} &\leq 2\frac{\rho\|(D-\mu I)^{-1}\|_2}{\rho\|(D-\mu I)^{-1}z\|_2^2}\sigma + \|(D-\mu I)^{-1}\|_2^{-1}\hat{\epsilon_1} + O(u_2\epsilon_1)\\ &\leq 2\|(D-\mu I)^{-1}z\|_2^{-1}\sigma + \|(D-\mu I)^{-1}\|_2^{-1}\epsilon_1 + O(u_2\epsilon_1). \qquad \square\end{aligned}$$

There are two terms in (20), one from the stopping criterion and one from rounding. For the stopping criterion to be effectively satisfied, the error from rounding should not be significantly larger.

Thus we would like

$$\sigma \approx \frac{\|(D-\mu I)^{-1}z\|_2}{\|(D-\mu I)^{-1}\|_2}\epsilon_1.$$

The precision u_2 is such that $\sigma = O(u_2)$. Hence a desirable condition is

$$u_2 \leq \frac{\|(D-\mu I)^{-1}z\|_2}{\|(D-\mu I)^{-1}\|_2}\epsilon_1.$$

A lower bound for $\|(D-\mu I)^{-1}z\|_2/\|(D-\mu I)^{-1}\|_2$ is $\min_{1\leq j\leq n}|\zeta_j|$, which leads to the condition

$$u_2 \leq \epsilon_1 \min_{1\leq j\leq n}|\zeta_j|.$$

In theory, this minimum could be zero, but we rule out zero components, since, in that case, the problem decouples. In general, we have a tolerance, call it ϵ_3, such that $|\zeta_j| \geq \epsilon_3$, $j = 1, 2, \ldots, n$. Then

$$u_2 \leq \epsilon_3\epsilon_1.$$

If $\epsilon_1, \epsilon_3 = O(u)$, where u is the standard machine unit, then $f(\mu)$ should be computed in double the current precision. The present single precision LAPACK routine [23] uses double precision for this part of the software. However, for the double precision routines, Sorensen and Tang [23] resolve this issue by using routines that simulate quadruple precision floating point operations in double precision.

Actually, the Sorensen–Tang procedure satisfies

$$\text{computed } (f(\mu)) = fl_2(\mu)(1+\hat{\sigma}),$$

where

$$\hat{\sigma} \leq \hat{c}u + O(u^2).$$

However, if we substitute that hypothesis into Theorem 3.4, the result holds with the change that

$$|\hat{\epsilon_1}| \leq \epsilon_1(1+(\hat{c}+n+5)u) + O(u^2\epsilon_1).$$

Remark (largely paraphrased from [23]). The routines in [23] work on any computer that satisfies the IEEE floating point standard. In fact, all that is necessary is that the floating point operations

$$A + B, \qquad A * B$$

produce exact results when the result is exactly representable. Also, for base 16 machines, it is necessary that these operations are always rounded to the nearest representable value.

CRAY's arithmetic is the one notable example of an arithmetic that does not satisfy these criteria. It satisfies the multiplication requirement, but not the addition requirement. Adding a guard digit in the addition/subtraction hardware would make CRAY's arithmetic satisfy all of the necessary assumptions.

For the above reason, as demonstrated recently by Kahan [18], *efficient* software for the solution of (6)–(7) on a CRAY is difficult, perhaps impossible, to develop. Software for this problem has been developed for CRAYs, but it is slow, even on a CRAY. The above discussion is considerably expanded in [23].

As a consequence of Theorem 3.4, from this point forward, we will make the assumption that (16) is satisfied.

4. Backward error analysis of updating procedures. In this section, we give a detailed error analysis of this updating procedure. The error analysis could be extended to any sequence of rank-one updates to a symmetric eigenvalue problem. As a result, we can give an analysis of the stability of Cuppen's procedure in floating point arithmetic. We will show that it is a stable method for finding all of the eigenvalues of a tridiagonal matrix.

We first cite the famous Newton–Kantorovich theorem. It will be used frequently in the arguments that follow. The theorem is stated for the sake of self-containment. Our statement of it is adapted from Dennis and Schnabel [7]. We consider the system of nonlinear equations

$$F(x) = 0, \qquad x \in \Re^n,$$

where $F(x) = (F_1(x), \ldots, F_n(x))^T$. In the statement of the theorem $\|\cdot\|$ is any vector norm or its induced matrix norm, depending upon context. For simplicity, define $N(x^{(0)}, r)$ as the set

$$N(x^{(0)}, r) = \{x : \|x - x^{(0)}\| < r\}.$$

LEMMA 4.1 (Kantorovich). *Let $r > 0$, $x^{(0)} \in \Re^n$, $F : \Re^n \rightarrow \Re^n$, and assume that F is continuously differentiable in the set $N(x^{(0)}, r)$. Let $J_F(x)$ be the Jacobian of F and assume that it satisfies the Lipschitz condition*

$$\|J_F(x^{(1)}) - J_F(x^{(2)})\| \leq \xi \|x^{(1)} - x^{(2)}\|$$

for all $x \in N(x^{(0)}, r)$. Assume also that $J_F(x^{(0)})$ is nonsingular and there exist constants β, η such that

$$\|J_F(x^{(0)})^{-1}\| \leq \beta, \qquad \|J_F(x^{(0)})^{-1} F(x^{(0)})\| \leq \eta.$$

Define $\xi_R = \beta\xi$; $\alpha = \xi_R \eta$. If $\alpha \leq \frac{1}{2}$ and $r \geq r_0 = (1 - \sqrt{1 - 2\alpha})/\beta\xi$ then the sequence $\{x^{(k)}\}$ produced by

$$x^{(k+1)} = x^{(k)} - J_F(x^{(k)})^{-1} F(x^{(k)}), \qquad k = 0, 1, \ldots$$

is well defined and converges to x^*, *a unique zero of* F *in* $N(x^{(0)}, r_0)$. *If* $\alpha < \frac{1}{2}$, *then* x^* *is the unique zero of* F *in* $N(x^{(0)}, r_1)$ *where* $r_1 = \min\{r, 1 + \sqrt{1 - 2\alpha}/\beta\xi\}$ *and*

$$\|x^* - x^{(k)}\| \le (2\alpha)^{2^k} \frac{\eta}{\alpha}, \qquad k = 0, 1, \ldots.$$

We really use the following result due to Dennis [6].

LEMMA 4.2. *Assume the hypothesis of Lemma* 4.1 *and* $\alpha < \frac{1}{2}$. *Then*

$$\tfrac{1}{2}\|x^{(k)} - x^{(k+1)}\| \le \|x^* - x^{(k)}\| \le 2\|x^{(k)} - x^{(k+1)}\|.$$

Better upper and lower bounds for the Newton iterates are given by Gragg and Tapia [11]. These are somewhat more complicated, but would slightly improve the bounds given below.

We will apply Lemma 4.2 to two different functions. The first is the spectral function $f(\lambda)$ in (7). The second function will be given in the proof of Theorem 4.8. We must verify that $r \ge r_0 = (1 - \sqrt{1 - 2\alpha})/\beta\xi$. Note that

$$r_0 = \frac{2\alpha}{(1 + \sqrt{1 - 2\alpha})\beta\xi} \le \frac{2\alpha}{\beta\xi} = 2\eta.$$

We need to have existence of the first and second derivatives in $N(\lambda^{(0)}, r)$. Here $r = 2\eta$, since that is the bound in Lemma 4.2. Note that

$$f'(\lambda) = \rho z^T (D - \lambda I)^{-2} z, \tag{24}$$

$$f''(\lambda) = 2\rho z^T (D - \lambda I)^{-3} z, \tag{25}$$

and that these derivatives exist unless $\lambda = \delta_i$ for some i. Thus for $\lambda \in (\delta_i, \delta_{i+1})$, we must have

$$\min\{|\lambda - \delta_i|, |\lambda - \delta_{i+1}|\} \ge 2\eta. \tag{26}$$

However, from (16),

$$\eta = \min\{|\lambda - \delta_i|, |\lambda - \delta_{i+1}|\}\epsilon_1. \tag{27}$$

From (27), (26) holds if $\epsilon_1 \le \frac{1}{2}$. Since $\epsilon_1 = O(u)$, this is a reasonable assumption. Thus, if the stopping criterion is satisfied, the spectral function is differentiable in a large enough interval.

Next we show that the constant α in Lemma 4.1 is less than $\frac{1}{2}$.

LEMMA 4.3. *Assume that* $\mu = \lambda_i^{(k)} \in (\delta_i, \delta_{i+1})$ *satisfies the stopping criterion* (16). *Let* $r = 2\eta$ *in Lemma* 4.1 *be defined by* (27). *Then* α *in Lemma* 4.1 *is bounded by*

$$\alpha \le 2\left(\frac{1}{1 - 2\epsilon_1}\right)^3 \epsilon_1.$$

Proof. Using the expressions (24)–(25) for $f'(\mu)$ and $f''(\mu)$ we have that

$$\alpha \le \max_{\omega \in [\mu - 2\eta, \mu + 2\eta]} \frac{|f''(\omega)|}{|f'(\mu)|} \min\{|\delta_i - \mu|, |\delta_{i+1} - \mu|\}\epsilon_1.$$

Let $y = (D - \mu I)^{-1}z$. Then

$$\begin{aligned}\alpha &\leq 2 \max_{\omega\in[\mu-2\eta,\mu+2\eta]} \frac{y^T(D-\omega I)^{-3}(D-\mu I)^2 y}{y^T y} \min\{|\delta_i - \mu|, |\delta_{i+1} - \mu|\}\epsilon_1 \\ &= 2 \max_{\omega\in[\mu-2\eta,\mu+2\eta]} \max_{1\leq j\leq n} \frac{(\delta_j - \mu)^2}{(\delta_j - \omega)^3} \min\{|\delta_i - \mu|, |\delta_{i+1} - \mu|\}\epsilon_1 \\ &\leq 2 \max_{\omega\in[\mu-2\eta,\mu+2\eta]} \max_{1\leq j\leq n} \left(\frac{\delta_j - \mu}{\delta_j - \omega}\right)^3 \epsilon_1 \\ &\leq 2 \max_{\omega\in[\mu-2\eta,\mu+2\eta]} \max_{1\leq j\leq n} \left(\frac{|\delta_j - \mu|}{|\delta_j - \mu| - |\mu - \omega|}\right)^3 \epsilon_1 \\ &\leq 2 \max_{1\leq j\leq n} \left(\frac{1}{1 - 2\eta/|\delta_j - \mu|}\right)^3 \epsilon_1.\end{aligned}$$

From (27), we have

$$\alpha \leq 2\left(\frac{1}{1-2\epsilon_1}\right)^3 \epsilon_1. \qquad \square$$

It should be noted that the condition

$$\alpha \leq \tfrac{1}{2}$$

from Lemma 4.1 will hold if

$$\epsilon_1 \leq 0.1145\ldots,$$

which is a root of the cubic equation

$$\tfrac{1}{2}(1 - 2\epsilon_1)^3 - 2\epsilon_1 = 0.$$

Again, $\epsilon_1 = O(u)$, so it is reasonable to assume that the hypothesis of Lemma 4.2 holds. Therefore, we may conclude that

$$(28) \qquad |\lambda_i^{(k)} - \lambda_i| \leq 2\min\{|\delta_i - \lambda_i^{(k)}|, |\delta_{i+1} - \lambda_i^{(k)}|\}\epsilon_1.$$

The following lemma from [8] shows us that we can also have nearly orthogonal eigenvectors.

LEMMA 4.4. *Suppose that $\hat{\lambda}$ and $\hat{\mu}$ are numerical approximations to the exact zeroes λ and μ of* (6). *Assume that the roots are distinct and let the relative errors for the quantities $\delta_i - \lambda$ and $\delta_i - \mu$ be denoted by θ_i and η_i, respectively. That is,*

$$\begin{aligned}\delta_i - \hat{\lambda} &= (\delta_i - \lambda)(1 + \theta_i), \\ \delta_i - \hat{\mu} &= (\delta_i - \mu)(1 + \eta_i)\end{aligned}$$

for $i = 1, 2, \ldots, n$. Let q_λ and q_μ be defined according to the formula (8) *using the computed quantities that satisfy* (16). *If $|\theta_i|, |\eta_i| \leq \epsilon$, then*

$$|q_{\hat{\lambda}}^T q_{\hat{\mu}}| \leq |q_\lambda^T E q_\mu| \leq \epsilon(2+\epsilon)\left(\frac{1+\epsilon}{1-\epsilon}\right)^2$$

with E a diagonal matrix whose ith diagonal element is

$$E_{ii} = \frac{\theta_i + \eta_i + \theta_i\eta_i}{(1+\theta_i)(1+\eta_i)} \left[\frac{f'(\lambda)f'(\mu)}{f'(\hat{\lambda})f'(\hat{\mu})} \right]^{\frac{1}{2}}.$$

The following corollary will be used extensively in our analysis of this procedure.

COROLLARY 4.5. *Assume that* (16) *is satisfied. Then*

$$|q_{\hat{\lambda}}^T q_{\hat{\mu}}| \leq 4\epsilon_1 + O(\epsilon_1^2),$$

where $\hat{\lambda}$ and $\hat{\mu}$ are distinct eigenvalues of $D + \rho zz^T$ that have been computed by (9)–(10).

The next lemma bounds the norm of the computed eigenvector matrix and its inverse.

LEMMA 4.6. *Assume that $\lambda_1, \ldots, \lambda_n$ satisfy the stopping criterion* (16) *and that*

$$\hat{Q} = (q_1, \ldots, q_n),$$

where $q_i = q_{\lambda_i}$ from Lemma 4.4. *Then*

$$\|\hat{Q}\|_2 \leq 1 + 2n\epsilon_1 + O(\epsilon_1^2), \qquad \|\hat{Q}^{-1}\|_2 \leq 1 + 2n\epsilon_1 + O(\epsilon_1^2). \tag{29}$$

Proof. From the bound (28),the hypothesis of Lemma 4.4 is satisfied with $\epsilon = 2\epsilon_1 + O(\epsilon_1^2)$. Thus from the conclusion, we have that

$$|q_i^T q_j| \leq 4\epsilon_1 + O(\epsilon_1^2).$$

Thus

$$\hat{Q}^T\hat{Q} = I + \delta Q,$$

where

$$\|\delta Q\|_F \leq 4n\epsilon_1 + O(\epsilon_1^2).$$

The conclusion on $\|\hat{Q}\|_2$ results from

$$\|\hat{Q}^T\hat{Q}\|_2 = \|Q\|_2^2 \leq 1 + 4n\epsilon_1 + O(\epsilon_1^2).$$

Thus $\|\hat{Q}\|_2$ satisfies (29). We still need to bound $\|\hat{Q}^{-1}\|_2$. By the Neumann lemma [10, p. 59], we have that

$$\begin{aligned} \|(\hat{Q}^T\hat{Q})^{-1}\|_2 = \|(I + \delta Q)^{-1}\|_2 \leq \frac{1}{1 - \|\delta Q\|_2} \leq \frac{1}{1 - 4n\epsilon_1} + O(\epsilon_1^2) \\ = 1 + 4n\epsilon_1 + O(\epsilon_1^2). \end{aligned}$$

Clearly,

$$\|Q^{-1}\|_2 = \sqrt{\|(\hat{Q}^T\hat{Q})^{-1}\|_2} \leq 1 + 2n\epsilon_1 + O(\epsilon_1^2). \qquad \square$$

We need the following technical lemma so that the Lipschitz condition in Lemma 4.1 can be established.

LEMMA 4.7. *Let*

$$J(z) = 2\operatorname{diag}(\gamma_1, \gamma_2, \dots, \gamma_n)V^T,$$

where

$$V = (v_1, v_2, \dots, v_n), \qquad v_i = (D - \lambda_i I)^{-1} z.$$

Then $J(z)$ satisfies

$$\|J(z^{(1)}) - J(z^{(2)})\|_2 \leq \xi \|z^{(1)} - z^{(2)}\|_2,$$

where

$$\xi = n^{1/2} \max_{1 \leq i \leq n} \gamma_i \|(D - \lambda_i I)^{-1}\|_2. \tag{30}$$

Proof. This is just a straightforward norm inequality. □

Lemmas 4.6 and 4.7 lead directly to a proof of Theorem 4.8, the main result of this paper. We show how to obtain the exact sequence of updates that yield the sequence of eigenvalues at each stage of the computation. This theorem considers the problem of finding the eigenvalues of

$$T = D + \sum_{j=1}^{s} \rho_j w_j w_j^T, \tag{31}$$

where $\|w_j\|_2 = 1, j = 1, 2, \dots, s$, using the procedure

$$D_0 = D, \tag{32}$$

$$D_j = Q_j^T (D_{j-1} + \rho_j z_j z_j^T) Q_j, \qquad k = 1, 2, \dots, s, \tag{33}$$

where $z_j = Q_{j-1} \dots Q_1 w_j$ and D_s is the diagonal matrix of the eigenvalues of T. The form (32)–(33) covers not only Cuppen's algorithm, but a variety of procedures.

THEOREM 4.8. *Assume that the method described by* (9)–(13) *is used with stopping criterion* (16) *to find the eigenvalues of T in* (31) *by the algorithm* (32)–(33). *Make the normalizing assumption that $\|w_j\|_2 = 1$, $j = 1, 2, \dots, s$. If for $j = 1, 2, \dots, s$ we have*

$$\tfrac{1}{2} n \epsilon_1 \|\hat{Q}_j^{-1}\|_2^2 \max_{1 \leq i \leq n} |\zeta_i^{(j)}|^{-1} < 1 \tag{34}$$

where $z_j = (\zeta_1^{(j)}, \dots, \zeta_n^{(j)})^T$, then the computed eigenvalues from this algorithm are the exact eigenvalues of

$$\hat{T} = D + \sum_{j=1}^{s} \rho_j \hat{w}_j \hat{w}_j^T, \tag{35}$$

where

$$\|\hat{w}_j - w_j\|_2 \leq n^{\frac{1}{2}} \epsilon_1 + O(\epsilon_1^2). \tag{36}$$

Proof. The proof is by induction on the number of updates, j. For $j = 0$, the theorem is trivial.

Assume true for $j-1$. Then

$$D_{j-1} = H_{j-1}^T \left(D_0 + \sum_{l=1}^{j-1} \rho_j \hat{w}_l \hat{w}_l^T \right) H_{j-1}$$

for some orthogonal matrix H_{j-1} (*which may not correspond to the computed eigenvector matrix*). Consider the application of the jth update in (33). Let λ_i, $i = 1, 2, \ldots, n$ be the computed eigenvalues at stage j. By the stopping criterion (16),

$$|f(\lambda_i)| \leq \frac{|f'(\lambda_i)|}{\|(D - \lambda_i I)^{-1}\|_2} \epsilon_1.$$

We note that $f'(\lambda_i) = \|(D - \lambda_i I)^{-1} z\|_2^2$, therefore,

$$|f(\lambda_i)| \leq \rho_j \frac{\|(D - \lambda_i I)^{-1} z_j\|_2^2}{\|(D - \lambda_i I)^{-1}\|_2} \epsilon_1 \leq \rho_i \|(D - \lambda_i I)^{-1} z_j\|_2 \epsilon_1 = \rho_j \gamma_i^{-1} \epsilon_1,$$

where

$$\gamma_i = \|(D - \lambda_i I)^{-1} z_j\|_2. \tag{37}$$

We now consider the system of nonlinear equations

$$g_i(\hat{z}_j) = \gamma_i \rho_j^{-1} \left(1 + \rho_j \sum_{l=1}^{n} \frac{\hat{\zeta}_l^2}{\delta_l - \lambda_i} \right) = 0, \qquad \hat{z}_j = (\hat{\zeta}_1, \ldots, \hat{\zeta}_n)^T. \tag{38}$$

Clearly, from (6)–(7), a vector $\hat{z}_j$ that is the root of (38) corresponds to the matrix $D + \rho \hat{z}_j \hat{z}_j^T$ that has exactly the eigenvalues $\lambda_1, \ldots, \lambda_n$. We note that the Jacobian of $g(z_j) = (g_1(z_j), \ldots, g_n(z_j))^T$ is $J(z_j)$, as given in Lemma 4.7. Thus $J(z_j)$ is Lipschitz continuous with Lipschitz constant ξ given by (30). Moreover, it is Lipschitz continuous for all values of z_j, thus we can satisfy the $r \geq r_0$ requirement of Lemma 4.1. We also have that $J(z_j) = 2Q_j$, thus the parameters β and η from Lemma 4.1 are

$$\|J(z_j)^{-1}\|_2 = \tfrac{1}{2} \|Q_j^{-1}\|_2 = \beta \leq \tfrac{1}{2} + n\epsilon_1 + O(\epsilon_1^2),$$

$$\|J(z_j)^{-1} g(z_k)\|_2 = \eta \leq \tfrac{1}{2} \|Q_j^{-1}\|_2 n^{\frac{1}{2}} \epsilon_1 \leq \tfrac{1}{2} n^{\frac{1}{2}} \epsilon_1 + O(\epsilon_1^2).$$

Therefore the condition $\alpha < \frac{1}{2}$ from Lemma 4.1 is satisfied if

$$\tfrac{1}{4} n \epsilon_1 \|Q_j^{-1}\|_2^2 \max_{1 \leq i \leq n} \gamma_i \|(D - \lambda_i I)^{-1}\|_2 < \tfrac{1}{2}.$$

From (37), we can conclude that

$$\begin{aligned} \gamma_i \|(D - \lambda_i I)^{-1}\|_2 &= \gamma_i \max\{\|(D - \lambda_i I)^{-1} e_i\|_2, \|(D - \lambda_i I)^{-1} e_{i+1}\|_2\} \\ &\leq \max\{|\zeta_i^{(j)}|^{-1}, |\zeta_{i+1}^{(j)}|^{-1}\}, \end{aligned}$$

hence (34) implies the hypothesis of Lemma 4.2.

From Lemma 4.2, we have

$$\|\hat{z}_j - z_j\|_2 \le \|Q_j^{-1}\|_2 n^{\frac{1}{2}}\epsilon_1.$$

Thus the $\hat{w}_k$ in (35)–(36) satisfies the bound

$$\|\hat{w}_j - w_j\|_2 \le \|Q_1 \cdots Q_j\|_2 \|\hat{z}_j - z_j\|_2 \le \|Q_1 \cdots Q_{j-1}\|_2 \|Q_j^{-1}\|_2 n^{\frac{1}{2}}\epsilon_1.$$

From Lemma 4.6, the above bound becomes

$$\|\hat{w}_j - w_j\|_2 \le (1+\nu)^{j+1} n^{\frac{1}{2}}\epsilon_1,$$

where $\nu = 2n\epsilon_1 + O(\epsilon_1^2)$. Hence (36). □

Again, note that we need a tolerance for the components of z_j, as discussed after Theorem 3.4 and by Sorensen and Tang [23].

A global bound on the perturbed eigenvalues from the Cuppen method is easily derived. The Wielandt–Hoffman theorem [24, pp. 104–105] states that if $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ are the eigenvalues of $\hat{T}$, and $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of T, then

$$\sum_{i=1}^{n} (\hat{\lambda}_i - \lambda_i)^2 \le \|\hat{T} - T\|_F^2.$$

That is used to obtain the following corollary.

COROLLARY 4.9. *Assume the hypothesis of Theorem* 4.8, *noting the assumption that* $\|w_j\|_2 = 1$, $j = 1, 2, \ldots, s$. *Let* $\hat{\lambda}_i, i = 1, 2, \ldots, n$ *be the computed eigenvalues of* T *in* (31) *by the algorithm* (32)–(33) *and let* $\lambda_i, i = 1, 2, \ldots, n$ *be the exact eigenvalues of* T. *Then*

$$\sqrt{\sum_{i=1}^{n} (\hat{\lambda}_i - \lambda_i)^2} \le 2n^{\frac{1}{2}} \|(\rho_1, \ldots, \rho_s)^T\|_2 \epsilon_1 + O(\epsilon_1^2). \tag{39}$$

Proof. From Theorem 4.8, the computed eigenvalues $\hat{\lambda}_i$, $i = 1, 2, \ldots, n$ are the exact eigenvalues of

$$\hat{T} = D_0 + \sum_{j=1}^{s} \rho_j \hat{w}_j \hat{w}_j^T,$$

where

$$\|\hat{w}_j - w_j\|_2 \le n^{\frac{1}{2}}\epsilon_1 + O(\epsilon_1^2).$$

The difference between $\hat{T}$ and T is bounded by

$$\|\hat{T} - T\|_F^2 \le \sum_{j=1}^{s} \rho_j^2 \|\hat{w}_j \hat{w}_j^T - w_j w_j^T\|_F^2. \tag{40}$$

Using the fact that

$$\begin{aligned} \|\hat{w}_j \hat{w}_j^T - w_j w_j^T\|_F &\le 2\|w_j\|_2 \|\hat{w}_j - w_j\|_2 + \|\hat{w}_j - w_j\|_2^2 \\ &\le 2n^{\frac{1}{2}}\epsilon_1 + O(\epsilon_1^2), \end{aligned}$$

and taking square roots in (40) yields (39). □

A second corollary yields a backward error bound for Cuppen's method.

COROLLARY 4.10. *Consider the algorithm in* §2 *for solving the symmetric tridiagonal eigenvalue problem. Let* $\hat{\lambda}_i$, $i = 1, 2, \ldots, n$ *be the computed eigenvalues and let* λ_i, $i = 1, 2, \ldots, n$ *be the exact eigenvalues. If we use the stopping criterion* (16), *then*

$$\sqrt{\sum_{i=1}^{n}(\hat{\lambda}_i - \lambda_i)^2} \leq 2\sqrt{2}n^{\frac{1}{2}}\|T\|_F\epsilon_1 + O(\epsilon_1^2).$$

Proof. The values of ρ_j, $j = 1, 2, \ldots, s$ are $\rho_j = \theta_j(1 + \theta_j^{-2})^{\frac{1}{2}}b_{k_j}$ where $k_j, j = 1, 2, \ldots, s$ are distinct values from the off-diagonals of T, and $\theta_j = \pm 1$. Thus $|\rho_j| = \sqrt{2}|b_{k_j}|$ so that

$$\|(\rho_1, \ldots, \rho_s)^T\|_2 = \sqrt{2}\|(b_{k_1}, \ldots, b_{k_s})^T\|_2 \leq \sqrt{2}\|T\|_F.$$

The application of Corollary 4.9 obtains the result. □

We note here that this result depends upon the choice $\theta = \pm 1$ and that the cancellation problem discussed by Dongarra and Sorensen [8] does not seem to be a factor in this analysis.

5. Applications to banded symmetric Toeplitz matrices. The results in this paper can also be directly applied to the analysis of a procedure due to Handy and Barlow [13] for finding the eigenvalues of banded, symmetric Toeplitz matrices. Numerical tests of this procedure are given in [13].

Define $\text{Toep}_n(a_1, a_2, \ldots, a_n)$ as the $n \times n$ symmetric matrix

$$\text{Toep}_n(a_1, a_2, \ldots, a_n) = \begin{pmatrix} a_1 & a_2 & \cdot & \cdot & \cdot & a_n \\ a_2 & a_1 & \cdot & \cdot & \cdot & a_{n-1} \\ \cdot & \cdot & a_1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_n & a_{n-1} & \cdot & \cdot & a_2 & a_1 \end{pmatrix}.$$

A Toeplitz matrix of bandwidth k is given by

$$\text{Toep}_n(a_1, a_2, \ldots, a_k, 0, \ldots, 0) = \begin{pmatrix} a_1 & a_2 & \cdot & a_k & 0 & 0 & \cdot \\ a_2 & a_1 & \cdot & \cdot & a_k & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_k & a_{k-1} & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & a_k & \cdot \\ 0 & 0 & \cdot & \cdot\cdot & a_{k-1} & a_k & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & a_k & \cdot & \cdot & a_2 & a_1 \end{pmatrix}.$$

We denote $\text{Toep}(a_1, \ldots, a_k, 0, \ldots, 0)$ as $T_{n,k}$. Bini and Capavoni [2] show that $T_{n,k}$ can be written as the sum of two matrices

$$T_{n,k} = C_{n,k} + H_{n,k-2}.$$

The second matrix $H_{n,k-2}$ is a Hankel matrix of the form

$$H_{n,k-2} = \begin{pmatrix} V_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & V_2 \end{pmatrix},$$

$$V_1 = \begin{pmatrix} a_3 & a_4 & \cdot & \cdot & a_k \\ a_4 & \cdot & \cdot & a_k & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_k & 0\cdot & \cdot & \cdot & 0 \end{pmatrix},$$

$$V_2 = \begin{pmatrix} 0 & \cdot & \cdot & 0 & a_k \\ 0 & \cdot & \cdot & a_k & a_{k-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_k & a_{k-1} & \cdot & \cdot & a_3 \end{pmatrix},$$

and the first matrix $C_{n,k}$ has the form

$$C_{n,k} = \text{Cross} - \text{Sum}_n(a_1 - a_3, a_2 - a_4, \ldots, a_{k-2} - a_k, a_{k-1}, a_k, 0, \ldots, 0),$$

where the notation above means that the first row is defined by the arguments and the subsequent rows are defined by the recurrence

$$(41) \qquad c_{i-1,j} + c_{i+1,j} = c_{i,j-1} + c_{i,j+1}, \qquad i,j = 1,2,\ldots,n$$

with the convention that $c_{0,j} = c_{i,0} = 0$. The matrix $C_{n,k}$ has a known eigenvalue decomposition

$$C_{n,k} = UDU^T,$$

where

$$(42) \qquad D = \text{diag}(\mu_1, \mu_2, \ldots, \mu_n), \quad U = (u_{ij}), \quad u_{ij} = \sqrt{\frac{2}{n+1}} \sin \frac{ij\pi}{n+1}.$$

The μ_i can be computed in kn multiplications using the relationship

$$\mu_j u_{ji} = \sum_{k=1}^{n} c_{ik} u_{jk},$$

and choosing a nonzero component u_{ij}. If we choose $i = \lceil \frac{n+1}{2j} \rceil$ then $|u_{ji}| \geq 1/\sqrt{n+1}$, thus the computed μ_i satisfy

$$(43) \qquad \sqrt{\sum_{i=1}^{n} (\tilde{\mu}_i - \mu_i)^2} \leq \sqrt{n+1} \|C_{n,k}\|_F u,$$

where u is the machine unit.

The eigenvalues of $T_{n,k}$ can be found using rank-one updates. In particular, the eigendecomposition of $H_{n,k-2}$ yields

$$H_{n,k-2} = \sum_{i=1}^{2k-2} \rho_i z_i z_i^T,$$

where $\rho_i, i = 1,2,\ldots,2k-4$ are the nonzero eigenvalues of $H_{n,k-2}$ and $z_i, i = 1,2,\ldots,2k-4$ are vectors with only $k-2$ nonzero components.

Thus the eigendecomposition of $T_{n,k}$ can be obtained from

$$(44) \qquad T_{n,k} = U(D + U^T H_{n,k-2} U) U^T$$

$$(45) \qquad = U \left(D + \sum_{i=1}^{2k-4} \rho_i w_i w_i^T \right) U^T, \quad w_i = U^T z_i, \quad i = 1,2,\ldots,n$$

$$(46) \qquad = U\hat{Q}\hat{D}\hat{Q}^T U^T.$$

The middle eigenproblem can be solved using $2k-4$ rank-one updates by the algorithm described in §2. The vectors $w_i = U^T z_i$ can be computed in $2(k-2)^2 n$ multiplications, or if $2k-2 > \log_2 n$ this can be reduced to $(k-2)n \log_2 n$ multiplications by using fast Fourier transforms.

By our theorems, the errors from the update operations satisfy the bounds

$$\sqrt{\sum_{i=1}^{n} (\tilde{\lambda}_i - \hat{\lambda}_i)^2} \leq \sqrt{n} \|(\rho_1, \ldots, \rho_{2k-4})^T\|_2 \epsilon_1 + O(\epsilon_1^2) \tag{47}$$

$$= \sqrt{n} \|H_{n,k-2}\|_F \epsilon_1 + O(\epsilon_1^2) \tag{48}$$

$$\leq \sqrt{n} \|T_{n,k}\|_F \epsilon_1 + O(\epsilon_1^2), \tag{49}$$

where $\hat{\lambda}_i$ are the true eigenvalues starting with

$$D = \mathrm{diag}(\tilde{\mu}_1, \ldots, \tilde{\mu}_n),$$

and $\tilde{\lambda}_i$, $i = 1, 2, \ldots, n$ are the computed eigenvalues. A straightforward argument yields the conclusion

$$\sqrt{\sum_{i=1}^{n} (\lambda_i - \tilde{\lambda}_i)^2} \leq (n^{\frac{1}{2}} \epsilon_1 + 2(n+1)^{\frac{1}{2}} u) \|T_{n,k}\|_F + O(u^2 + \epsilon_1^2). \tag{50}$$

Hence, this procedure is numerically stable. Handy and Barlow [13] show that this approach can be faster than band QR when $k \ll n$.

6. Conclusions.

We have shown rigorous backward error bounds on eigenvalue update algorithms. These bounds show that the method of Cuppen [4] with slight alterations to the implementation of Dongarra and Sorensen [8] and Sorensen and Tang [23] is stable in the classical sense [24], [20], [10]. However, it is not known whether the algorithm satisfies the stronger stability criterion described by Barlow and Demmel [1].

These results confirm those in the tests by Jessup and Ipsen [14], [17], [15] and those in the tests by Handy and Barlow [13], [12].

Acknowledgments. The author is very grateful to Jim Bunch and the referees. Their suggestions greatly improved this paper. He also thanks Jim Demmel, Danny Sorensen, and Peter Tang for their help.

REFERENCES

[1] J. L. Barlow and J. W. Demmel, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.

[2] D. Bini and M. Capovani, *Spectral and computational properties of band symmetric Toeplitz matrices*, Linear Algebra Appl., 52/53 (1983), pp. 98–126.

[3] J. R. Bunch, C. P. Nielsen, and D. C. Sorensen, *Rank-one modification of the symmetric eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.

[4] J. Cuppen, *A divide and conquer method for the symmetric tridiagonal eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.

[5] J. W. Demmel, J. J. Dongarra, J. DuCroz, A. Greenbaum, S. Hammarling, and D. Sorensen, *A prospectus for the development of a linear algebra library for high-performance computers*, Tech. Rep. MCS-TM-97, Mathematics and Computer Science Div., Argonne National Laboratory, Argonne, IL, Sept. 1987.

[6] J. E. Dennis, Jr., *Toward a unified convergence theory for Newton-like methods*, in Nonlinear Function Analysis and Applications, L. B. Rall, ed., Academic Press, New York, 1971, pp. 425–472.

[7] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall, Englewood Cliffs, NJ, 1983.

[8] J. J. Dongarra and D. C. Sorensen, *A fully parallel algorithm for the symmetric eigenproblem*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 139–154.

[9] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–344.

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

[11] W. B. Gragg and R. A. Tapia, *Optimal error bounds for the Newton–Kantorovich theorem*, SIAM J. Numer. Anal., 11 (1974), pp. 10–13.

[12] S. L. Handy, *On the Numerical Solution of the Eigenvalue and Inverse Eigenvalue Problem for Real, Symmetric Toeplitz Matrices*, Ph.D. thesis, Dept. of Computer Science, The Pennsylvania State Univ., University Park, PA, 1990.

[13] S. L. Handy and J. L. Barlow, *The numerical solution of banded, Toeplitz eigenvalue problems*, Tech. Rep. CS-92-27, Dept. of Computer Science, The Pennsylvania State Univ., University Park, PA, December 1992; SIAM J. Matrix Anal. Appl., 1994, to appear.

[14] I. C. F. Ipsen and E. R. Jessup, *Solving the symmetric tridiagonal eigenvalue problem on the hypercube*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 203–229.

[15] E. R. Jessup, *Parallel Solution of the Symmetric Tridiagonal Eigenproblem*, Ph.D. thesis, Dept. of Computer Science, Yale University, New Haven, CT, 1989.

[16] E. R. Jessup and D. C. Sorensen, *A parallel algorithm for computing the singular value decomposition*, Tech. Rep. MCS-TM-102, Mathematics and Computer Science Div., Argonne National Laboratory, Argonne, IL, 1987.

[17] E. R. Jessup and I. C. F. Ipsen, *Improving the accuacy of inverse iteration*, Res. Rep. YALEU/DCS/RR-767, Dept. of Computer Science, Yale Univ., New Haven, CT, March 1990; SIAM J. Sci. Statist. Comput., 13 (1992), pp. 550–572.

[18] W. Kahan, *On rank one modifications to the symmetric eigenproblem*, Plenary Lecture, Householder Symposium X, Tylösand, Sweden, June 1990.

[19] J. J. Moré, *The Levenberg–Marquardt algorithm: Implementation and theory*, in Proc. Dundee Conf. Numerical Analysis, G. A. Watson, ed., Springer-Verlag, New York, 1978.

[20] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, NJ, 1980.

[21] C. H. Reinsch, *Smoothing by spline functions*, Numer. Math., 10 (1967), pp. 177–183.

[22] ———, *Smoothing by spline functions* II, Numer. Math., 16 (1971), pp. 451–454.

[23] D. C. Sorensen and P. T. Tang, *On the orthogonality of eigenvectors computed by the divide-and-conquer techniques*, Tech. Rep. MCS-P152-0490, Mathematics and Computer Science Div., Argonne National Laboratory, Argonne, IL, 1990.

[24] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 619–628, July 1993

JACOBI'S METHOD FOR SKEW-SYMMETRIC MATRICES*

DEREK HACON†

Abstract. A Jacobi-type method is given for reducing real skew-symmetric matrices to real Schur form. 4×4 rotations are used to reduce the "off-diagonal" part of the matrix. Asymptotic quadratic convergence is established for the special cyclic method, provided the eigenvalues are distinct.

Key words. Jacobi, quaternions

AMS subject classification. 65F

Introduction. Jacobi's method for orthogonally diagonalizing symmetric matrices is based upon the fact that any symmetric 2×2 matrix S may be diagonalized by means of a rotation matrix which is expressible in terms of S by means of a simple formula.

In this article we present a Jacobi-type method for orthogonally reducing a real skew-symmetric matrix to real Schur form, a question first considered in [P]. Evidently, in the skew-symmetric case, 2×2 rotations are of no use. Our method uses 4×4 rotations to annihilate off-diagonal 2×2 submatrices. It is possible to use 3×3 rotations instead but the resulting algorithm is more complicated.

Our main result is Theorem 1, which asserts that, given a 4×4 skew-symmetric matrix A, there exists a simple formula for calculating a rotation matrix which will transform A into real Schur form. An essential ingredient of the formula is the (well-known) parametrization of 4×4 rotations by pairs of quaternions. Despite the presence of the latter, the resulting algorithm is fairly simple, which may make it attractive in the context of parallel computing.

1. 4×4 matrices and quaternions. In this section we review some of the basic properties of quaternions and how these are related to the algebra of real 4×4 matrices.

Real 4×4 matrices form a 16-dimensional (normed) algebra. The norm is the Frobenius norm

$$\|A\| = \sqrt{\sum_{k,l} a_{kl}^2}. \tag{1.1}$$

The quaternions are a four-dimensional vector space over the real numbers with the standard basis $\{1, i, j, k\}$. The multiplication rules

$$i^2 = j^2 = k^2 = ijk = -1 \tag{1.2}$$

make the quaternions an associative algebra. The typical quaternion $a + bi + cj + dk$ is denoted by q. The real part of q is a and the pure quaternion part is $bi + cj + dk$. $\bar{q}$, the conjugate of q, is given by

$$\bar{q} = a - bi - cj - dk. \tag{1.3}$$

The modulus of q is the Euclidean norm of q, namely,

$$|q| = \sqrt{a^2 + b^2 + c^2 + d^2}. \tag{1.4}$$

* Received by the editors May 8, 1990; accepted for publication (in revised form) October 23, 1991.

† PUC-Rio, Departamento de Matemática, Rua Marquês de São Vicente, 225, Gávea, 22453, Rio de Janeiro, Brazil. The work of this author was supported in part by FINEP and CNPq, Brazil (haco@LNCC.BITNET).

An easy calculation shows that

$$q\bar{q} = a^2 + b^2 + c^2 + d^2. \tag{1.5}$$

From the last formula we have

$$|q|^2 = q\bar{q} = |\bar{q}|^2; \tag{1.6}$$

conjugation is related to multiplication by

$$\overline{pq} = \bar{q}\bar{p}. \tag{1.7}$$

Now $q\bar{q}$ is real and therefore

$$|pq|^2 = pq\overline{pq} = pq\bar{q}\bar{p} = p\bar{p}q\bar{q}. \tag{1.8}$$

Hence

$$|pq| = |p|\,|q|. \tag{1.9}$$

From the formula $q\bar{q} = |q|^2$ we deduce the distinguishing feature of the quaternions which is that any nonzero quaternion has a multiplicative inverse. The formula for q^{-1} is

$$q^{-1} = \frac{\bar{q}}{|q|^2}. \tag{1.10}$$

For any pair of quaternions p and q, we define $p*q$ to be the matrix (relative to the basis 1, i, j, k) of the linear transformation which sends a quaternion v to $pv\bar{q}$. Thus $p*q$ is a real 4×4 matrix and such matrices multiply together according to the rule

$$(p*q)(r*s) = (pr)*(qs). \tag{1.11}$$

If $q = a + bi + cj + dk$, then

$$1*q = \begin{pmatrix} a & b & c & d \\ -b & a & -d & c \\ -c & d & a & -b \\ -d & -c & b & a \end{pmatrix} \quad\text{and}\quad q*1 = \begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix}. \tag{1.12}$$

The 16 matrices

$$\frac{1*1}{2}, \frac{i*1}{2}, \frac{1*i}{2}, \ldots, \frac{k*k}{2}$$

are, in fact, an orthonormal basis (relative to the Frobenius norm) for the 4×4 matrices. We shall not use this fact.

2. 4 × 4 rotations. By analogy with two- and three-dimensional rotations, we use the word rotation to mean "orthogonal linear transformation of determinant +1."

LEMMA 1. *If p and q are quaternions of modulus one, then $p*q$ is a rotation.*

Proof. Both $p*1$ and $1*q$ are easily seen to be orthogonal matrices. Furthermore, a quick calculation shows that $p*1$ and $1*q$ both have positive determinants. Thus both $p*1$ and $1*q$ (and therefore also $p*q$) are rotations.

We note that if $p = q$, then the rotation $p*q$ sends 1 to 1 and is therefore a rotation of the 3-space of pure quaternions.

LEMMA 2. *If π is a pure quaternion of modulus one, then $\pi * \pi$ is a* 180° *rotation of the pure quaternions, and its axis of rotation passes through π.*

Proof. Since $\pi\pi\bar{\pi} = \pi$ the axis of rotation of $\pi * \pi$ passes through π. Furthermore, $(\pi * \pi)^2 = \pi^2 * \pi^2 = -1 * -1 = 1 * 1$. Hence $\pi * \pi$ is a notation through 180° or 360°. It cannot be through 360° because, clearly, $\pi * \pi \neq 1 * 1$.

LEMMA 3. *If θ and ϕ are pure quaternions of modulus one, then the rotation*

$$\frac{1-\theta\phi}{|1-\theta\phi|} * \frac{1-\theta\phi}{|1-\theta\phi|} \tag{2.1}$$

takes ϕ into θ.

Proof. Every rotation of 3-space may be written as the product of two 180° rotations. To rotate ϕ to θ, we first rotate through 180° about the bisector of the angle $\theta O \phi$. This takes ϕ to θ and γ to $-\gamma$, where γ is any pure quaternion perpendicular to θ and ϕ. Next we rotate through 180° about the axis through θ. This leaves θ where it is and brings $-\gamma$ back to γ. Since the bisector of angle $\theta O \phi$ passes through $\theta + \phi$ the required rotation is

$$(\theta * \theta)\left(\frac{\theta+\phi}{|\theta+\phi|} * \frac{\theta+\phi}{|\theta+\phi|}\right), \tag{2.2}$$

which equals

$$\frac{\theta\phi-1}{|\theta\phi-1|} * \frac{\theta\phi-1}{|\theta\phi-1|},$$

which, in turn, is clearly the same thing as

$$\frac{1-\theta\phi}{|1-\theta\phi|} * \frac{1-\theta\phi}{|1-\theta\phi|}.$$

The only time this formula does not work is when $\theta + \phi = 0$.

3. Reduction of 4 × 4 skew-symmetric matrices.

We can now write down our formula for the rotation which reduces a given 4 × 4 skew-symmetric matrix to real Schur form.

We will need the following characterization of skew-symmetric matrices.

LEMMA 4. *The* 4 × 4 *matrix A is skew-symmetric if and only if*

$$A = \pi * 1 + 1 * \psi, \tag{3.1}$$

where π and ψ are pure quaternions. Furthermore, $\pi * 1 + 1 * \psi$ is in real Schur form if and only if both π and ψ are real multiples of i.

Proof. This follows immediately from the matrix expressions for $\pi * 1$ and $1 * \psi$. Writing

$$\pi = ui + vj + wk \quad \text{and} \quad \psi = xi + yj + zk, \tag{3.2}$$

we have

$$\pi * 1 + 1 * \psi = \begin{pmatrix} 0 & x-u & y-v & z-w \\ u-x & 0 & -w-z & v+y \\ v-y & w+z & 0 & -u-x \\ w-z & -v-y & u+x & 0 \end{pmatrix}, \tag{3.3}$$

from which π and ψ may be found in terms of A.

THEOREM 1. *If A is the skew-symmetric matrix $\pi * 1 + 1 * \psi$ and R the rotation matrix*

$$\frac{|\pi| - i\pi}{||\pi| - i\pi|} * \frac{|\psi| - i\psi}{||\psi| - i\psi|}, \tag{3.4}$$

then RAR^{-1} is in real Schur form.

Proof. We seek p and q, both of modulus one, such that

$$(p*q)(\pi*1 + 1*\psi)(p*q)^{-1} = si*1 + 1*ti, \tag{3.5}$$

where s and t are real.

In other words, we want

$$p\pi p^{-1} = si \quad \text{and} \quad q\psi q^{-1} = ti. \tag{3.6}$$

By Lemma 3 we should choose

$$p = \frac{|\pi| - i\pi}{|\text{numerator}|} \quad \text{and} \quad q = \frac{|\psi| - i\psi}{|\text{numerator}|}. \tag{3.7}$$

Then $R = p*q$ is a rotation matrix and

$$RAR^{-1} = |\pi|i*1 + 1*|\psi|i. \tag{3.8}$$

This formula can be checked by direct calculation using the equation

$$(|\pi| - i\pi)\pi = |\pi|i(|\pi| - i\pi), \tag{3.9}$$

and similarly for ψ.

If we prefer, we can replace i by $-i$ in the above formula. We choose the sign according to the following rule, which is the analogue of the condition $|\theta| \leqq \frac{\pi}{4}$ in the symmetric case.

Let $\pi = ui + vj + wk$. Choose

$$p = \frac{|\pi| - i\pi}{|\text{numerator}|} \quad \text{if } u \geqq 0$$

and

$$p = \frac{|\pi| + i\pi}{|\text{numerator}|} \quad \text{if } u < 0, \tag{3.10}$$

and similarly for ψ. The exceptional case $\pi = 0$ is dealt with by setting $p = 1$, and similarly for $\psi = 0$.

4. Asymptotic quadratic convergence. By analogy with the symmetric case, one may define the various Jacobi methods (cyclic, etc.). We need only deal with the $2n \times 2n$ case because the $2n + 1 \times 2n + 1$ case can easily be reduced to this, as follows. Let $\tilde{A}$ be a $2n + 1 \times 2n + 1$ skew-symmetric matrix. As its determinant is zero, there exists a vector (v_1, say) of norm one such that $\tilde{A}v_1 = 0$. Completing to an orthonormal basis $v_1, v_2, \ldots, v_{2n+1}$, we get an orthogonal matrix (G, say) such that

$$G\tilde{A}G^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & A \end{pmatrix}. \tag{4.1}$$

A is, of course, also skew-symmetric, so we are reduced to the $2n \times 2n$ case. An obvious question now is how fast the algorithm converges. We content ourselves with considering the special cyclic method. In this case, adapting an argument of Wilkinson [W], we are able to establish asymptotically quadratic convergence, provided the eigenvalues are distinct. Sadly, the proof of the general case, given in [K], does not carry over (the fly in the ointment being the equation $A_1^{(k)} - \lambda_1 I =$ etc, in [K, p. 20]). We deal with the row-cyclic method, the proof for the column-cyclic method being similar.

By the phrase "distinct eigenvalues," we mean that the eigenvalues of the skew-symmetric matrix A (necessarily pure imaginary) are $\lambda_1 i, -\lambda_1 i, \ldots, \lambda_n i, -\lambda_n i$, where

$$0 \leqq \lambda_1 < \lambda_2 < \cdots < \lambda_n. \tag{4.2}$$

We define

$$\delta = \min_{r \neq s} |\lambda_r - \lambda_s|. \tag{4.3}$$

A may be partitioned into 2×2 blocks:

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \vdots & \vdots \\ A_{n1} & \cdots & A_{nn} \end{pmatrix}, \tag{4.4}$$

$$\text{Define Off}(A) = \begin{pmatrix} O & A_{12} & \cdots & A_{1n} \\ A_{21} & O & & \\ \vdots & & \ddots & \\ A_{n1} & & & O \end{pmatrix}; \tag{4.5}$$

in other words, Off(A) is A with each A_{ii} replaced by O.

THEOREM 2. *Let A be a skew-symmetric matrix with distinct eigenvalues (as defined above). Suppose one sweep of the special cyclic Jacobi method transforms A into the matrix Z. Then, provided that $\|\text{Off}(A)\| < \delta$, we will have*

$$\|\text{Off}(Z)\| \leqq \frac{\|\text{Off}(A)\|^2}{\sqrt{2}\varepsilon}, \tag{4.6}$$

where $\varepsilon = \delta - \|\text{Off}(A)\|$.

Proof. We will need the following key inequality.

LEMMA 5. *Suppose the first rotation of the sweep (i.e., the one that annihilates A_{12} and A_{21}) transforms A into B. Then, for $m \neq 1, 2$,*

$$\|B_{1m}\| \leqq \|A_{1m}\| + \frac{\|A_{12}\| \|A_{2m}\|}{\varepsilon}. \tag{4.7}$$

Proof. Let

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \pi * 1 + 1 * \psi.$$

Writing $\pi = ui + vj + wk$ and $\psi = xi + yj + zk$, we have, as in (3.3),

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} 0 & x-u & y-v & z-w \\ u-x & 0 & -w-z & v+y \\ v-y & w+z & 0 & -u-x \\ w-z & -v-y & u+x & 0 \end{pmatrix}. \tag{4.8}$$

Assume that

$$u \geqq 0 \quad \text{and} \quad x \geqq 0. \tag{4.9}$$

The first rotation of the sweep is $R = p*q$, where

$$p = \frac{|\pi| + u + wj - vk}{|\text{numerator}|} \tag{4.10}$$

and

$$q = \frac{|\psi| + x + zj - yk}{|\text{numerator}|}\,. \tag{4.11}$$

Setting $Q = (\begin{smallmatrix} R & O \\ O & I \end{smallmatrix})$, we have $B = QAQ^{-1}$, where $B_{12} = 0 = B_{21}$. For $m \neq 1, 2$,

$$\begin{pmatrix} B_{1m} \\ B_{2m} \end{pmatrix} = R \begin{pmatrix} A_{1m} \\ A_{2m} \end{pmatrix} \tag{4.12}$$

and $B_{lm} = A_{lm}$ if both $l \neq 1, 2$ and $m \neq 1, 2$. Thus, for $m \neq 1, 2$, one has

$$\sum_i \|B_{im}\|^2 = \sum_i \|A_{im}\|^2. \tag{4.13}$$

$\|\text{Off}(B)\| \leqq \|\text{Off}(A)\|$ because

$$\|\text{Off}(B)\|^2 = \|\text{Off}(A)\|^2 - 2\|A_{12}\|^2. \tag{4.14}$$

Each 2×2 block of both $p*1$ and $1*q$ is a scalar multiple of an orthogonal matrix. For a 2×2 matrix W, let us write

$$|W| = \frac{1}{\sqrt{2}} \|W\|. \tag{4.15}$$

If W and M are 2×2 matrices and W is a scalar multiple of an orthogonal matrix, we have

$$\|WM\| = |W|\,\|M\|, \tag{4.16}$$

because W is $|W|$ times an orthogonal matrix.

Let us write

$$p*1 = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \quad \text{and} \quad 1*q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}. \tag{4.17}$$

For $m \neq 1, 2$ the first rotation of the sweep transforms

$$\begin{pmatrix} A_{1m} \\ A_{2m} \end{pmatrix}$$

into

$$\begin{pmatrix} B_{1m} \\ B_{2m} \end{pmatrix} = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} A_{1m} \\ A_{2m} \end{pmatrix}. \tag{4.18}$$

Thus

$$B_{1m} = (P_{11}Q_{11} + P_{12}Q_{21})A_{1m} + (P_{11}Q_{12} + P_{12}Q_{22})A_{2m}, \tag{4.19}$$

so that

$$\|B_{1m}\| \leqq |P_{11}|\,|Q_{11}|\,\|A_{1m}\| + \cdots + |P_{12}|\,|Q_{22}|\,\|A_{2m}\|. \tag{4.20}$$

Now $|P_{11}|^2 + |P_{12}|^2 = 1$ since $|p| = 1$, and

$$|Q_{11}|^2 + |Q_{21}|^2 = 1 \quad \text{since } |q| = 1. \tag{4.21}$$

Thus $|P_{11}|\,|Q_{11}| + |P_{12}|\,|Q_{21}| \leqq 1$. Hence

$$\|B_{1m}\| \leqq \|A_{1m}\| + (|P_{12}| + |Q_{12}|)\|A_{2m}\|. \tag{4.22}$$

From the formulas for p and q

$$|P_{12}| \leqq \frac{\sqrt{v^2 + w^2}}{|\pi| + u} \quad \text{and} \quad |Q_{12}| \leqq \frac{\sqrt{y^2 + z^2}}{|\psi| + x}. \tag{4.23}$$

Therefore,

$$|P_{12}| + |Q_{12}| \leqq \frac{\sqrt{v^2 + w^2}}{2u} + \frac{\sqrt{y^2 + z^2}}{2x}. \tag{4.24}$$

To show that x and u are not too small, we apply the Wielandt–Hoffmann Theorem to $A = (A - \mathrm{Off}(A)) + \mathrm{Off}(A)$. We assume, in addition to $u \geqq 0$, $x \geqq 0$, that $u \geqq x \geqq 0$. We deduce, in particular, that there are eigenvalues αi, βi of A (where α and β are, of course, real) such that, for δ as in (4.3),

$$\beta - \delta \geqq \alpha \geqq 0$$

and

$$(u - x - \alpha)^2 + (u + x - \beta)^2 \leqq \tfrac{1}{2}\|\mathrm{Off}(A)\|^2. \tag{4.25}$$

Interpreting this in terms of distances in the plane, the distance between (α, β) and $(u - x, u + x)$ is at most

$$\frac{\|\mathrm{Off}\, A\|}{\sqrt{2}}.$$

Projecting onto the line $x + y = 0$, it follows that

$$\frac{|2x|}{\sqrt{2}} \geqq \frac{|\beta - \alpha|}{\sqrt{2}} - \frac{\|\mathrm{Off}\, A\|}{\sqrt{2}} \tag{4.26}$$

and therefore that

$$x \geqq \frac{\varepsilon}{2}.$$

As we assumed $u \geqq x$, we also have

$$u \geqq \frac{\varepsilon}{2}. \tag{4.27}$$

If $x \geqq u \geqq 0$, we deduce likewise that $u \geqq \frac{\varepsilon}{2}$. So

$$\begin{aligned} \|B_{1m}\| &\leqq \|A_{1m}\| + \frac{\sqrt{v^2 + w^2} + \sqrt{y^2 + z^2}}{\varepsilon}\|A_{2m}\| \\ &\leqq \|A_{1m}\| + \frac{\sqrt{2}}{\varepsilon}\left(\sqrt{v^2 + w^2 + y^2 + z^2}\right)\|A_{2m}\|, \end{aligned} \tag{4.28}$$

but

$$(4.29) \qquad \|A_{12}\|^2 = 2(v^2 + w^2 + y^2 + z^2).$$

Therefore,

$$(4.30) \qquad \|B_{1m}\| \leqq \|A_{1m}\| + \frac{\|A_{12}\|\,\|A_{2m}\|}{\varepsilon},$$

as required.

If $u < 0$, (4.10) is replaced by

$$(4.31) \qquad p = \frac{|\pi| - u - wj + vk}{|\text{numerator}|},$$

and similarly for $x < 0$. In (4.23) and (4.24) u and x are replaced by $|u|$ and $|x|$ and we now conclude that $|u| \geqq \frac{\varepsilon}{2}$ and $|x| \geqq \frac{\varepsilon}{2}$, and that therefore Lemma 5 holds, whatever the signs of u and x may be. We observe that for $\varepsilon > 0$ the exceptional cases $\pi = 0$ and $\psi = 0$, cannot occur because neither u nor x can be zero. □

With this inequality at our disposal, we may now prove Theorem 2 along the lines of [W]. As the sweep proceeds, δ, of course, remains unchanged and $\|\text{Off}\,A\|$ does not increase, so we may use the same ε in the analogous inequality at each stage of the sweep.

We will prove (4.6) for $n = 5$. The proof for arbitrary n is similar. The sweep transforms A into Z as follows.

(4.32)

$$A = \begin{pmatrix} S_0 & A_0 & B_0 & C_0 & D_0 \\ \cdot & T_0 & E_0 & F_0 & G_0 \\ \cdot & \cdot & U_0 & H_0 & I_0 \\ \cdot & \cdot & \cdot & V_0 & J_0 \\ \cdot & \cdot & \cdot & \cdot & W_0 \end{pmatrix} \to \begin{pmatrix} S_1 & O & B_1 & C_1 & D_1 \\ \cdot & T_1 & E_1 & F_1 & G_1 \\ \cdot & \cdot & U_0 & H_0 & I_0 \\ \cdot & \cdot & \cdot & V_0 & J_0 \\ \cdot & \cdot & \cdot & \cdot & W_0 \end{pmatrix}$$

$$\to \begin{pmatrix} S_2 & A_2 & O & C_2 & D_2 \\ \cdot & T_1 & E_2 & F_1 & G_1 \\ \cdot & \cdot & U_1 & H_1 & I_1 \\ \cdot & \cdot & \cdot & V_0 & J_0 \\ \cdot & \cdot & \cdot & \cdot & W_0 \end{pmatrix} \to \begin{pmatrix} S_3 & A_3 & B_3 & O & D_3 \\ \cdot & T_1 & E_2 & F_2 & G_1 \\ \cdot & \cdot & U_1 & H_2 & I_1 \\ \cdot & \cdot & \cdot & V_1 & J_1 \\ \cdot & \cdot & \cdot & \cdot & W_0 \end{pmatrix}$$

$$\to \begin{pmatrix} S_4 & A_4 & B_4 & C_4 & O \\ \cdot & T_1 & E_2 & F_2 & G_2 \\ \cdot & \cdot & U_1 & H_2 & I_2 \\ \cdot & \cdot & \cdot & V_1 & J_2 \\ \cdot & \cdot & \cdot & \cdot & W_1 \end{pmatrix} \to \begin{pmatrix} S_4 & A_5 & B_5 & C_4 & O \\ \cdot & T_2 & O & F_3 & G_3 \\ \cdot & \cdot & U_2 & H_3 & I_3 \\ \cdot & \cdot & \cdot & V_1 & J_2 \\ \cdot & \cdot & \cdot & \cdot & W_1 \end{pmatrix}$$

$$\to \begin{pmatrix} S_4 & A_6 & B_5 & C_5 & O \\ \cdot & T_3 & E_4 & O & G_4 \\ \cdot & \cdot & U_2 & H_4 & I_3 \\ \cdot & \cdot & \cdot & V_2 & J_3 \\ \cdot & \cdot & \cdot & \cdot & W_1 \end{pmatrix} \to \begin{pmatrix} S_4 & A_7 & B_5 & C_5 & D_5 \\ \cdot & T_4 & E_5 & F_5 & O \\ \cdot & \cdot & U_2 & H_4 & I_4 \\ \cdot & \cdot & \cdot & V_2 & J_4 \\ \cdot & \cdot & \cdot & \cdot & W_2 \end{pmatrix}$$

$$\to \begin{pmatrix} S_4 & A_7 & B_6 & C_6 & D_5 \\ \cdot & T_4 & E_6 & F_6 & O \\ \cdot & \cdot & U_3 & O & I_5 \\ \cdot & \cdot & \cdot & V_3 & J_5 \\ \cdot & \cdot & \cdot & \cdot & W_2 \end{pmatrix} \to \begin{pmatrix} S_4 & A_7 & B_7 & C_6 & D_6 \\ \cdot & T_4 & E_7 & F_6 & G_5 \\ \cdot & \cdot & U_4 & H_6 & O \\ \cdot & \cdot & \cdot & V_3 & J_6 \\ \cdot & \cdot & \cdot & \cdot & W_3 \end{pmatrix}$$

$$\to \begin{pmatrix} S_4 & A_7 & B_7 & C_7 & D_7 \\ \cdot & T_4 & E_7 & F_7 & G_7 \\ \cdot & \cdot & U_4 & H_7 & I_7 \\ \cdot & \cdot & \cdot & V_4 & J_7 \\ \cdot & \cdot & \cdot & \cdot & W_4 \end{pmatrix} = Z.$$

(The dots stand for $-A_0^T$ and so on.)

Applying (4.30) three times, we have

$$(4.33) \qquad \|A_4\| \leqq \frac{\|B_1\|}{\varepsilon}\|E_1\| + \frac{\|C_2\|}{\varepsilon}\|F_1\| + \frac{\|D_3\|}{\varepsilon}\|G_1\|.$$

Similarly,

$$(4.34) \qquad \|B_4\| \leqq \frac{\|C_2\|}{\varepsilon}\|H_1\| + \frac{\|D_3\|}{\varepsilon}\|I_1\|$$

and

$$\|C_4\| \leqq \frac{\|D_3\|}{\varepsilon}\|J_1\|.$$

Hence

(4.35)

$$\begin{aligned} \|A_4\|^2 + \|B_4\|^2 + \|C_4\|^2 &\leqq \frac{1}{\varepsilon^2}(\|B_1\|^2 + \|C_2\|^2 + \|D_3\|^2)(\|E_1\|^2 + \|F_1\|^2 + \|G_1\|^2) \\ &\quad + \frac{1}{\varepsilon^2}(\|C_2\|^2 + \|D_3\|^2)(\|H_1\|^2 + \|I_1\|^2) \\ &\quad + \frac{1}{\varepsilon^2}(\|D_3\|^2)(\|J_1\|^2) \\ &\leqq \frac{1}{\varepsilon^2}(\|B_1\|^2 + \|C_2\|^2 + \|D_3\|^2)(\|E_1\|^2 + \|F_1\|^2 + \|G_1\|^2 \\ &\quad + \|H_1\|^2 + \|I_1\|^2 + \|J_1\|^2). \end{aligned}$$

The argument now proceeds just as in [W], mutatis mutandis. At the end of the sweep we will have

$$(4.36) \qquad \frac{1}{2}\|\mathrm{Off}(Z)\|^2 \leqq \frac{\|\mathrm{Off}(A)\|^2}{2\varepsilon^2} \cdot \frac{\|\mathrm{Off}(A)\|^2}{2},$$

which proves Theorem 2.

Thus, for distinct eigenvalues, the convergence of the special cyclic method is asymptotically quadratic.

As to convergence, we observe that the classical Jacobi method can be shown to converge in the usual way. If, in A, we annihilate the off-diagonal A_{ij} of largest norm, we obtain B (say) where

$$\|\mathrm{Off}(B)\| \leqq \sqrt{1 - \frac{2}{n^2 - n}}\,\|\mathrm{Off}(A)\|,$$

and this guarantees that the off-diagonal norm tends to zero.

Acknowledgment. The author is grateful to Manchester University, Manchester, England, for its hospitality.

REFERENCES

[K] H. P. M. Van Kempen, *On the quadratic convergence of the special cyclic Jacobi method*, Numer. Math., 9 (1966), pp. 19–22.

[P] M. H. C. Paardekooper, *An eigenvalue algorithm for skew symmetric matrices*, Numer. Math., 17 (1971), pp. 189–202.

[W] J. H. Wilkinson, *Note on the quadratic convergence of the cyclic Jacobi process*, Numer. Math., 4 (1962), pp. 296–300.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 629-645, July 1993

ON THE BEZOUTIAN STRUCTURE OF THE MOORE-PENROSE INVERSES OF HANKEL MATRICES*

GEORG HEINIG† AND FRANK HELLINGER†

Abstract. It is shown that the Moore-Penrose inverses of Hankel matrices are generalized Bezoutians. This result generalizes the well-known fact that Hankel matrix inverses are (classical) Bezoutians.

Key words. Hankel matrix, Toeplitz matrix, generalized inverse, Moore–Penrose inverse, Bezoutian matrix

AMS subject classifications. 15A09, 15A57, 47B35

1. Introduction. In the present paper, we investigate the structure of the Moore–Penrose inverse (MPI) of Hankel matrices

$$H = \begin{bmatrix} s_0 & s_1 \cdots s_{n-1} \\ s_1 & s_2 \cdots s_n \\ \vdots & \vdots \\ s_{m-1} & s_m \cdots s_{m+n-2} \end{bmatrix} \tag{1.1}$$

with complex entries. The main aim is to extend some well-known results on the structure of the inverses of Hankel matrices to MPIs. These results open the door to the development of fast algorithms for the construction of the Moore–Penrose inverses of Hankel matrices and related matrix classes.

A convenient tool to describe the results, both the classical ones and those of the present paper, is the *generating function of a matrix defined as follows*: If $A = [a_{ij}]_0^{m-1}{}_0^{n-1}$, then $A(\lambda, \mu)$ denotes the polynomial in two variables

$$A(\lambda, \mu) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} a_{ij}\lambda^i\mu^j.$$

DEFINITION 1.1. An $n \times n$ matrix B is said to be a *classical* (*Hankel*) *Bezoutian* if and only if it has a generating function of the form

$$B(\lambda, \mu) = \frac{a(\lambda)b(\mu) - b(\lambda)a(\mu)}{\lambda - \mu}, \tag{1.2}$$

where $a(\lambda)$ and $b(\lambda)$ are polynomials with degree less than or equal to n. The matrix B is called the *Bezoutian of the two polynomials* $a(\lambda)$ *and* $b(\lambda)$ and is denoted by Bez (a, b).

The Bezoutian concept was introduced in connection with root localization problems (problems of the Routh–Hurwitz type). For information about the history of this concept, we refer the reader to [19] and [25].

It is well known that the inverse of a nonsingular Hankel matrix is a classical Bezoutian. For the positive definite case, this fact already follows from the Christoffel–Darboux formula of the theory of orthogonal polynomials; in the general case it is a consequence of the Gohberg–Semencul formula (see [8] and also [13], [18], and [21]) for the inverse of a nonsingular Toeplitz matrix. This fact was first observed by Lander

* Received by the editors December 10, 1990; accepted for publication (in revised form) October 22, 1991.

† Universität Leipzig, Fachbereich Mathematik Informatik, Augustusplatz, Leipzig D-O-7010, Germany (heide@mathematik.uni-leipzig.dbp.de or ghein@hadrian.hrz.su-chemnitz.de).

in [20]. In this paper, it is also proved that, vice versa, the inverse of a regular Bezoutian is a Hankel matrix.

If the Hankel matrix H is singular, then the question is whether there is a generalized inverse of H, which is a Bezoutian. By a generalized inverse of a matrix A, we mean a matrix B satisfying

$$(1.3) \qquad ABA = A, \qquad BAB = B.$$

In [13, § I, 7.2], it is shown that any $n \times n$ Hankel matrix possesses a generalized inverse that is a classical Bezoutian.

A singular matrix has many generalized inverses. The most important of them is MPI, denoted by A^+, which fulfills the following additional conditions:

$$(1.4) \qquad (A^+A)^* = A^+A, \qquad (AA^+)^* = AA^+.$$

Here A^* denotes the conjugate transpose of A. The MPI of a matrix exists always and is uniquely determined (see [4]).

One can easily see that the MPI of a Hankel matrix is not necessarily a classical Bezoutian. As an example, we consider the 3×3 Hankel matrix

$$H = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

The MPI of this matrix equals

$$H^+ = \frac{1}{4}\begin{bmatrix} 1 & 0 & 1 \\ 0 & 4 & 0 \\ 1 & 0 & 1 \end{bmatrix},$$

which is not a classical Bezoutian.

In order to describe the structure of the MPI of Hankel matrices, we introduce the concept of an r-Bezoutian.

DEFINITION 1.2. An $m \times n$ matrix B is said to be a (*Hankel*) *r-Bezoutian* if and only if there are polynomials $u_i(\lambda)$ and $v_i(\lambda)$ ($i = 1, \dots, r$) such that

$$(1.5) \qquad (\lambda - \mu)B(\lambda, \mu) = \sum_{i=1}^{r} u_i(\lambda)v_i(\mu).$$

The minimal r for which a representation (1.5) exists will be called the *Bezoutian rank* of B.

For a given matrix B, let ∇B denote the matrix with the generating function $(\lambda - \mu)B(\lambda, \mu)$. If $B = [b_{ij}]$, then ∇B is given by

$$\nabla B = [b_{i-1,j} - b_{i,j-1}].$$

It is an immediate consequence of the definition that the Bezoutian rank of B equals the rank of the matrix ∇B.

Clearly, classical Bezoutians are special 2-Bezoutians. It can be shown (see [13, p. 36]) that any nonsingular 2-Bezoutian is classical.

Furthermore, it is obvious that any symmetric 2-Bezoutian is classical. With the help of symplectic linear algebra, this can be generalized to the following assertion: If the Bezoutian rank of a symmetric matrix B is equal to r then r is even and B can be represented as the sum of $r/2$ classical Bezoutians.

In fact, let $\nabla B = GF^T$ be a full rank decomposition (G, $F \in \mathbb{C}^{n \times r}$). Since B is symmetric, ∇B is skew-symmetric. This implies the existence of a nonsingular skew-symmetric $r \times r$ matrix C satisfying $G = FC$. It is well known that the order of any nonsingular skew-symmetric matrix C is even, and C can be represented in the form

$$C = K\begin{bmatrix} 0 & I_m \\ -I_m & 0 \end{bmatrix}K^T,$$

where $K \in \mathbb{C}^{r \times r}$. From this representation, our claim follows immediately.

For $r = 2$ the r-Bezoutian concept was introduced in the more general block matrix context by Anderson and Jury [2]. The general r-Bezoutian concept appears for the first time in the papers of Lerer and Tismenetsky [21], [22] (see also [9] and [24]).

Let us note that there is a close relationship between the concept of an r-Bezoutian and the concept of matrices with a displacement structure, which was introduced and extensively studied by Kailath, Kung, and Morf [17] and Kailath and Chun [18]. Actually the (Hankel) displacement rank of an $n \times n$ matrix B equals the rank of a certain $n \times n$ submatrix of the $(n+1) \times (n+1)$ matrix ∇B. Hence the displacement rank is less than or equal to the Bezoutian rank. More discussion about these topics can be found in [3], [6]–[8], [19], and references therein.

Knowing from the example presented above that the MPI of a Hankel matrix is in general not a 2-Bezoutian, it is natural to ask whether it is an r-Bezoutian for certain r. The positive answer of this fact will be the contents of the present paper. We are going to formulate the main results.

THEOREM 1.1. *The* MPI *of a Hankel matrix is a* 4-*Bezoutian.*

In the square case the result can still be strengthened.

THEOREM 1.2. *The* MPI *of a square Hankel matrix is the sum of two classical Bezoutians.*

The proof of these theorems will be presented in § 3. In § 6 the exact Bezoutian rank of H^+ for the square case $m = n$ is computed. In order to formulate the corresponding theorem, we introduce the concept of a Hankel circulant.

DEFINITION 1.3. An $n \times n$ matrix H is said to be an *α-Hankel circulant* ($\alpha \in \mathbb{C}$) if and only if it is Hankel, $H = [s_{i+j}]$, and

$$s_{i+n} = \alpha s_i \qquad (i = 0, \dots, n-2).$$

We call H an *∞-Hankel circulant* if and only if

$$s_i = 0 \qquad (i = 0, \dots, n-2).$$

When speaking of a Hankel circulant, we have in mind an α-circulant for $\alpha \in \mathbb{C}$ or $\alpha = \infty$.

After reversing the order of rows or columns, a Hankel circulant goes over into a (generalized) circulant in the usual sense (see [5] or § 4 for the definition). For $\alpha = 1$ a Hankel circulant is a g-circulant in the sense of [5]. It is well known that the Hankel circulants are the only Hankel matrices for which the inverse is again a Hankel matrix (see [10], [16], and [23]).

THEOREM 1.3. *Let H be a nonzero square Hankel matrix and let r denote the Bezoutian rank of H^+. Then*

(a) $r = 2$ *if H is regular or a Hankel circulant;*

(b) $r = 4$ *otherwise.*

In addition to this theorem, let us note that in the nonsquare case the description of the Bezoutian rank is more complicated. In particular, there are Hankel matrices the MPI of which has Bezoutian rank 3. As an example, we put

$$H = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

The MPI of this matrix is given by

$$H^{+} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{2} \\ 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix},$$

and the Bezoutian rank of H^{+} equals 3.

In [5] it is proved in principle that the MPI of a 1- or (-1)-Hankel circulant is a Hankel circulant again. This result immediately extends to the case $|\alpha| = 1$. However, it fails to be true if $|\alpha| \neq 1$. As an example we choose a matrix of the form

$$H = \begin{bmatrix} H_0 & 0 \\ 0 & 0 \end{bmatrix},$$

where H_0 is a regular upper triangular Hankel matrix. H is a 0-Hankel Bezoutian. The MPI of H is given by

$$H^{+} = \begin{bmatrix} H_0^{-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Clearly, H^{+} is not a Hankel circulant.

From Theorem 1.3, it follows now, however, that the MPI of a Hankel circulant is a 2-Hankel Bezoutian. Since it is symmetric, it is actually a classical Bezoutian.

The fact that the MPI of Hankel matrices are Bezoutians is not only of theoretical interest, but has important computational consequences. Let us explain this fact.

In § 4 it will be shown that r-Bezoutians can be represented as the sum of r products of circulant matrices. The formulas are, in a sense, generalizations of the Gohberg–Semencul formula and the formulas of Ammar and Gader [1]. For more details we refer to [11] (see also [18] and [21]). The multiplication of a circulant matrix by a vector is equivalent to the cyclic convolution of two vectors. That means, in order to compute the vector By for an $n \times n$ r-Bezoutian B, one can apply fast convolution algorithms or fast Fourier transform (FFT), which means that the computational complexity can be reduced to $O(n \log n)$ floating point operations (flops). Thus the pseudo-solution of a system $Hx = y$, which is $x = H^{+}y$, can be computed with this amount, provided that the parameters of the matrix representations are given.

Let us shortly explain from where these parameters come. In order to evaluate the MPI of a Hankel matrix, we utilize the familiar extension approach, which means we extend H to a regular matrix $\mathscr{H}$. To get a convenient extension, we apply a kernel structure theorem for Hankel matrices. This leads to a matrix $\mathscr{H}$, which is a Hankel mosaic matrix in the sense of [14]. Now we can apply a result of [14], which is related to the block generalization of the Gohberg–Semencul formula (see [7], [18], and [21]) and states that the inverse of $\mathscr{H}$ can be constructed from a system of eight vectors via a Bezoutian-type formula. The vector system is called a *fundamental system*. It also determines the MPI H^{+}. In § 5 we present various possibilities to characterize and construct a fundamental system.

Of course, now the problem arises of how to compute the fundamental system effectively, i.e., via a fast algorithm with complexity less than $O(n^3)$. This problem was treated in Hellinger's diploma thesis [15] and will be considered in a subsequent paper of the authors in this journal. Actually, it can be shown that $O(n^2)$ complexity algorithms to evaluate fundamental systems do exist.

Toeplitz matrices $[a_{i-j}]$ are closely related to Hankel matrices. A Toeplitz matrix goes over in a Hankel matrix after changing the order of columns or rows. In that way all results formulated in this paper can immediately be transformed into the corresponding Toeplitz matrix results. The Hankel Bezoutian concept has to be replaced by the following one.

DEFINITION 1.4. A matrix A is said to be an *r-Toeplitz Bezoutian* if and only if there are polynomials $u_i(\lambda)$, $v_i(\lambda)$ $(i = 1, \dots, r)$ such that

$$(1 - \lambda\mu)A(\lambda, \mu) = \sum_{i=1}^{r} u_i(\lambda)v_i(\mu).$$

2. The extension approach for the MPI. In order to construct the MPI, we use the familiar extension approach, which is based on the following property.

LEMMA 2.1. [4]. *Let A be an arbitrary matrix, U a matrix the columns of which form a basis of the kernel of A, and V a matrix the columns of which form a basis of the kernel of A^*. Then the matrix*

$$\mathscr{A} = \begin{bmatrix} A & V \\ U^* & 0 \end{bmatrix} \tag{2.1}$$

is regular and $\mathscr{A}^{-1}$ admits the representation

$$\mathscr{A}^{-1} = \begin{bmatrix} A^+ & (U^*)^+ \\ V^+ & 0 \end{bmatrix},$$

where the superscript "+" *denotes the* MPI.

This lemma indicates that in order to construct the MPI of a matrix A, one has to find the kernels of A and A^*. For Hankel matrices there exists a remarkable kernel description. This description concerns more a family of Hankel matrices than a single matrix H. For given complex numbers s_i $(i = 0, \dots, m + n - 2)$, we consider the family of Hankel matrices H_k $(k = 1, \dots, m + n - 1)$,

$$H_k = [s_{i+j}]_{i=0}^{l-1}{}_{j=0}^{k-1} \quad (l = m + n - k). \tag{2.2}$$

We formulate the corresponding result in polynomial language. If $x = (x_k)_0^{n-1}$, then we denote $x(\lambda) = x_0 + x_1\lambda + \cdots + x_{n-1}\lambda^{n-1}$. Instead of $\ker H_k$, we describe the polynomial subspaces

$$\mathscr{C}_k = \{u(\lambda) \colon H_k u = 0\}.$$

THEOREM 2.1 [13]. *Let $\{H_k\}$ be a family of Hankel matrices given by* (2.2), *$H_k \neq 0$. Then there exist integers p and q, $p \leqq q$, with $p + q = m + n$ and vectors $u \in \mathbb{C}^{p+1}$, $v \in \mathbb{C}^{q+1}$ such that the following is true*:

(a) *If $p < k \leqq q$, then $\{u(\lambda), \dots, \lambda^{k-p-1}u(\lambda)\}$ is a basis of $\mathscr{C}_k$*;

(b) *If $q < k$, then $\{u(\lambda), \dots, \lambda^{k-p-1}u(\lambda), v(\lambda), \dots, \lambda^{k-q-1}v(\lambda)\}$ forms a basis of $\mathscr{C}_k$*;

(c) *If $p \geqq k$, then $\mathscr{C}_k = \{0\}$*.

In order to construct the MPI, we also need the kernel of the adjoint matrices H_k^*. But we have $H_k^* = \bar{H}_l$, where $l = m + n - k$ and the bar denotes the matrix with conjugate complex entries. Hence for $\mathscr{C}_k^* := \{x(\lambda) \colon x \in \ker H_k^*\}$ we have $\mathscr{C}_k^* = \bar{\mathscr{C}}_l$.

Now we consider the matrix H defined by (1.1), which is equal to H_n. The integers

$$\varkappa := n - p, \qquad \nu := n - q \quad (= p - m)$$

are called *partial indices of* H. A system of vectors $\{u, v\}$ or corresponding polynomials $\{u(\lambda), v(\lambda)\}$ possessing the properties formulated in Theorem 2.1 is called a *fundamental system* of H. A fundamental system of a Hankel matrix can be constructed recursively with $O((m+n)^2)$ operations or, if FFT is utilized, with $O((m+n)\log^2(m+n))$ operations (see [12] and references therein).

For a given vector $u = (u_k)_0^p$, let $M_n(u)$ $(n > p)$ denote the $n \times (n-p)$ Hankel matrix

$$M_n(u) = \left.\begin{bmatrix} 0 & & & u_0 \\ & & u_0 & u_1 \\ & \cdot^{\cdot^{\cdot}} & & \vdots \\ u_0 & & \cdots & u_p \\ \vdots & & \cdot^{\cdot^{\cdot}} & \\ u_{p-1} & u_p & & \\ u_p & & 0 & \end{bmatrix}\right\} n.$$

According to Theorem 2.1, we distinguish between cases (a) $\varkappa \geqq 0 \geqq \nu$, (b) $\nu > 0$, and (c) $\varkappa < 0$.

THEOREM 2.2. *Let H be an $m \times n$ nonzero Hankel matrix, $\{u, v\}$ a fundamental system, and $\varkappa$, ν the partial indices of H, $\varkappa \geqq \nu$.*

(a) *If $\varkappa \geqq 0 \geqq \nu$, then*

$$\mathscr{H} := \begin{bmatrix} H & M_m(\bar{u}) \\ M_n(\bar{u})^T & 0 \end{bmatrix}\text{[1]} \tag{2.3}$$

is nonsingular and

$$\mathscr{H}^{-1} = \begin{bmatrix} H^+ & (M_n(\bar{u})^T)^+ \\ M_m(\bar{u})^+ & 0 \end{bmatrix}. \tag{2.4}$$

(b) *If $\nu > 0$, then*

$$\mathscr{H} := \begin{bmatrix} H \\ M_n(\bar{u})^T \\ M_n(\bar{v})^T \end{bmatrix}$$

is nonsingular and

$$\mathscr{H}^{-1} = [H^+ \quad W^+], \quad \textit{where } W = \begin{bmatrix} M_n(\bar{u})^T \\ M_n(\bar{v})^T \end{bmatrix}. \tag{2.5}$$

(c) *If $\varkappa < 0$, then*

$$\mathscr{H} := [H \quad M_m(\bar{u}) \quad M_m(\bar{v})]$$

[1] If $\varkappa = 0$ or $\nu = 0$, then $M_m(\bar{u})$ or $M_n(\bar{u})^T$ does not appear in $\mathscr{H}$.

is nonsingular and

$$\mathscr{H}^{-1} = \begin{bmatrix} H^+ \\ W^+ \end{bmatrix}, \quad \textit{where } W = [M_m(\bar{u}) \quad M_m(\bar{v})].$$

3. Hankel mosaic matrices and proof of Theorem 1.1. In the previous section, the problem of the MPI of Hankel matrices H is reduced to the problem of the inversion of a matrix $\mathscr{H}$ consisting of Hankel blocks. Such a matrix is referred to as a Hankel mosaic matrix. Matrices of this type were investigated in [14]. We present some results of the paper [14], which will be used for the proof of the main theorems.

Let $\mathscr{H} = [H_{ij}]_{i=1}^{r}{}_{j=1}^{s}$ be a nonsingular Hankel mosaic matrix, where the H_{ij} are $m_i \times n_j$ Hankel blocks. Furthermore, let $\mathscr{B} = [B_{ji}]_{j=1}^{s}{}_{i=1}^{r}$ be the inverse of $\mathscr{H}$, where the B_{ji} are $n_j \times m_i$ blocks.

For notation purposes, we need a block generalization of the generating function concept defined in § 1. For two multi-indices $N = (n_1, \dots, n_s)$ and $M = (m_1, \dots, m_r)$, let $\mathbb{C}^{N \times M}$ denote the space of all block matrices $A = [A_{ij}]_{i=1}^{s}{}_{j=1}^{r}$, where $A_{ij} \in \mathbb{C}^{n_i \times m_j}$. Now $A(\lambda, \mu)$ is defined as the $s \times r$ polynomial matrix

$$[A_{ij}(\lambda, \mu)]_{i=1}^{s}{}_{j=1}^{r},$$

where $A_{ij}(\lambda, \mu)$ means the generating function in the usual sense. If, in particular, all m_j equal 1, then $\mathbb{C}^{N \times M}$ will be the space of block columns $X = \operatorname{col}(X_i)_1^s$, $X_i \in \mathbb{C}^{n_i \times M}$, and the generating function is a polynomial only in λ.

We still need some more notation. The last unit vector $(0, \dots, 0, 1)$ in $\mathbb{C}^m$ will be denoted by $e(m)$. If $M = (m_1, \dots, m_r)$, then $e(M)$ will be the block diagonal matrix $\operatorname{diag}(e(m_1), \dots, e(m_r)) \in \mathbb{C}^{M \times \mathbf{1}(r)}$, where $\mathbf{1}(r) := (1, \dots, 1)$ (r times). For a given Hankel matrix $H = [h_{k+l}]_{k=0}^{m-1}{}_{l=0}^{n-1}$, we introduce column vectors

$$g(k, H) := [h_{n+m-k} \cdots h_{n+m-2} h_{n+m-1}]^T \in \mathbb{C}^k,$$

where h_{n+m-1} stands for an arbitrary fixed number. For a Hankel mosaic matrix $\mathscr{H} = [H_{ij}]_{i=1}^{r}{}_{j=1}^{s}$ we define

$$g(M, \mathscr{H}) := [g(m_i, H_{ij})]_{i=1}^{r}{}_{j=1}^{s} \in \mathbb{C}^{M \times \mathbf{1}(s)}.$$

THEOREM 3.1 [14]. *Let* $\mathscr{H} = [H_{ij}]_1^r{}_1^s \in \mathbb{C}^{M \times N}$, $M = \{m_1, \dots, m_r\}$, $N = \{n_1, \dots, n_s\}$, *be a nonsingular Hankel mosaic matrix. If* $X \in \mathbb{C}^{N \times \mathbf{1}(r)}$, $W \in \mathbb{C}^{N \times \mathbf{1}(s)}$, $\tilde{X} \in \mathbb{C}^{M \times \mathbf{1}(s)}$, $\tilde{W} \in \mathbb{C}^{M \times \mathbf{1}(r)}$, *are the solutions of the equations*

$$\begin{aligned} &(3.1) \qquad & \mathscr{H}X = e(M), \qquad & \mathscr{H}^T\tilde{X} = e(N), \\ & & \mathscr{H}W = g(M, \mathscr{H}), \qquad & \mathscr{H}^T\tilde{W} = g(N, \mathscr{H}^T), \end{aligned}$$

respectively, then the generating function of $\mathscr{H}^{-1} \in \mathbb{C}^{N \times M}$ *is given by*

$$(3.2) \qquad \mathscr{H}^{-1}(\lambda, \mu) = (X(\lambda)\tilde{Y}(\mu)^T - Y(\lambda)\tilde{X}(\mu)^T)/(\lambda - \mu),$$

where

$$Y(\lambda) = W(\lambda) - e(N + \mathbf{1}(s))(\lambda), \qquad \tilde{Y}(\lambda) = \tilde{W}(\lambda) - e(M + \mathbf{1}(r))(\lambda).$$

Now we are able to prove Theorem 1.1. We distinguish between the cases (a), (b), and (c) of Theorem 2.2.

Proof of Theorem 1.1. *Case* (a). In this case we have $r = s = 2$ and $X(\lambda)$, $Y(\lambda)$, $\tilde{X}(\lambda)$, and $\tilde{Y}(\lambda)$ admit representations

$$X(\lambda) = \begin{bmatrix} x_1(\lambda) & x_2(\lambda) \\ \xi_1(\lambda) & \xi_2(\lambda) \end{bmatrix}, \qquad Y(\lambda) = \begin{bmatrix} y_1(\lambda) & y_2(\lambda) \\ \eta_1(\lambda) & \eta_2(\lambda) \end{bmatrix},$$

$$\tilde{X}(\lambda) = \begin{bmatrix} \tilde{x}_1(\lambda) & \tilde{x}_2(\lambda) \\ \tilde{\xi}_1(\lambda) & \tilde{\xi}_2(\lambda) \end{bmatrix}, \qquad \tilde{Y}(\lambda) = \begin{bmatrix} \tilde{y}_1(\lambda) & \tilde{y}_2(\lambda) \\ \tilde{\eta}_1(\lambda) & \tilde{\eta}_2(\lambda) \end{bmatrix},$$

where $x_i, y_i \in \mathbb{C}^n$, $\xi_i, \eta_i \in \mathbb{C}^{-\nu}$, $\tilde{x}_i, \tilde{y}_i \in \mathbb{C}^m$, $\tilde{\xi}_i, \tilde{\eta}_i \in \mathbb{C}^{\varkappa}$. Applying part (a) of Theorem 2.2, we get

$$(3.3) \qquad H^+(\lambda, \mu) = \frac{1}{\lambda - \mu}\,(x_1(\lambda)\,\tilde{y}_1(\mu) + x_2(\lambda)\,\tilde{y}_2(\mu) - y_1(\lambda)\,\tilde{x}_1(\mu) - y_2(\lambda)\,\tilde{x}_2(\mu)).$$

That means H^+ is a 4-Bezoutian.

Case (b). In this case $r = 3$ and $s = 1$. Therefore, we have representations

$$X(\lambda) = [x_1(\lambda)x_2(\lambda)x_3(\lambda)], \qquad Y(\lambda) = y_1(\lambda),$$

$$\tilde{X}(\lambda) = [\tilde{x}_1(\lambda)\tilde{x}_2(\lambda)\tilde{x}_3(\lambda)]^T, \qquad \tilde{Y} = \begin{bmatrix} \tilde{y}_1(\lambda) & \tilde{y}_2(\lambda) & \tilde{y}_3(\lambda) \\ \tilde{\eta}_1(\lambda) & \tilde{\eta}_2(\lambda) & \tilde{\eta}_3(\lambda) \\ \tilde{\sigma}_1(\lambda) & \tilde{\sigma}_2(\lambda) & \tilde{\sigma}_3(\lambda) \end{bmatrix}.$$

Applying Theorem 2.2, we obtain

$$(3.4) \qquad H^+(\lambda, \mu) = \frac{1}{\lambda - \mu}\,(x_1(\lambda)\,\tilde{y}_1(\mu) + x_2(\lambda)\,\tilde{y}_2(\mu) + x_3(\lambda)\,\tilde{y}_3(\mu) - y_1(\lambda)\,\tilde{x}_1(\lambda)),$$

which shows that H^+ is a 4-Bezoutian also in this case.

Case (c). Case (c) is analogous to (b). Thus, Theorem 1.1 is proved. □

Proof of Theorem 1.2. In the case of a square matrix H, the matrix $\mathscr{H}$ is symmetric and $g(M, \mathscr{H}) = g(N, \mathscr{H}^T)$. Hence $X = \tilde{X}$ and $W = \tilde{W}$, which implies $x_i = \tilde{x}_i$ and $y_i = \tilde{y}_i$ for $i = 1, 2$. Thus we get from (3.3)

$$H^+ = \operatorname{Bez}(x_1, y_1) + \operatorname{Bez}(x_2, y_2),$$

which proves Theorem 1.2. □

4. Matrix representations. Let A be an r-Bezoutian. We show that A can be represented as a product sum of circulant matrices, in particular, of triangular Toeplitz and Hankel matrices. Without loss of generality we may assume that A is square. Otherwise we would add some zero columns or rows to get a square matrix.

For a given vector $g = (g_i)_0^n$ and complex α, let $C_\alpha(g)$ denote the so-called (α-Toeplitz) circulant matrix

$$C_\alpha(g) = \begin{bmatrix} g_0 + \alpha g_n & g_1 & \cdots & g_{n-1} \\ \alpha g_{n-1} & g_0 + \alpha g_n & \cdots & g_{n-2} \\ \vdots & & \ddots & \vdots \\ \alpha g_1 & \alpha g_2 & & g_0 + \alpha g_n \end{bmatrix}.$$

We have

$$(4.1) \qquad C_\alpha(g) = g(Z_\alpha),$$

where Z_α denotes the matrix

$$Z_\alpha = \begin{bmatrix} 0 & 1 & & 0 \\ & & \ddots & \\ & \ddots & & \\ 0 & & & 1 \\ \alpha & 0 & & 0 \end{bmatrix}.$$

We consider matrices of the form

$$B = \sum_{i=1}^{r} C_\alpha(g_i)^T C_\beta(h_i), \tag{4.2}$$

where $g_i, h_i \in \mathbb{C}^{n+1}$ and $\alpha, \beta \in \mathbb{C}$.

LEMMA 4.1. *Let B be of the form* (4.2). *Assume that for all complex $\lambda \neq 0$*

$$\sum_{i=1}^{r} g_i(\lambda^{-1}) h_i(\lambda) = 0. \tag{4.3}$$

Then

$$B(\lambda, \mu) = \frac{1 - \alpha\beta}{1 - \lambda\mu} \sum_{i=1}^{r} g_i(\lambda) h_i(\mu). \tag{4.4}$$

Proof. We introduce the vectors $l(\lambda) = [1\ \lambda \cdots \lambda^{n-1}]^T$. Then $B(\lambda, \mu) = l(\lambda)^T B l(\mu)$. Furthermore,

$$Z_\beta l(\mu) \equiv \mu l(\mu) \qquad \mathrm{mod}\ (\mu^n - \beta)$$

and

$$l(\lambda)^T Z_\alpha^T \equiv \lambda l(\lambda)^T \qquad \mathrm{mod}\ (\lambda^n - \alpha).$$

Hence we get, taking (4.1) into account,

$$\begin{aligned} B(\lambda, \mu) &= \sum_{i=1}^{r} l(\lambda)^T g_i(Z_\alpha)^T h_i(Z_\beta) l(\mu) \\ &\equiv \sum_{i=1}^{r} g_i(\lambda) h_i(\mu) l(\lambda)^T l(\mu) \qquad \mathrm{mod}\ (\lambda^n - \alpha, \mu^n - \beta) \\ &\equiv \sum_{i=1}^{r} g_i(\lambda) h_i(\mu) \frac{1 - \lambda^n \mu^n}{1 - \lambda\mu} \qquad \mathrm{mod}\ (\lambda^n - \alpha, \mu^n - \beta). \end{aligned}$$

In view of (4.3), $\sum g_i(\lambda) h_i(\mu)$ is divisible by $1 - \lambda\mu$. Hence we may employ the relations $\lambda^n \equiv \alpha$ mod $(\lambda^n - \alpha)$ and $\mu^n \equiv \beta$ mod $(\mu^n - \beta)$ and obtain (4.4) as a congruence. Since both sides of (4.4) have degree less than n in λ and μ, the congruence is actually an identity. □

Let J denote the matrix of the flip operator

$$J = \begin{bmatrix} 0 & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & 0 \end{bmatrix}.$$

LEMMA 4.2. *Suppose that $A = JB$ and $(1 - \lambda\mu)B(\lambda, \mu) = b(\lambda, \mu)$. Then $(\lambda - \mu)A(\lambda, \mu) = a(\lambda, \mu)$, where $a(\lambda, \mu) = b(\lambda^{-1}, \mu)\lambda^n$. In particular, A is an r-Hankel Bezoutian if and only if B is an r-Toeplitz Bezoutian.*

Proof. We have $l(\lambda)^T J = \lambda^{n-1} l(\lambda^{-1})^T$ for all $\lambda \neq 0$. Hence

$$\begin{aligned} A(\lambda, \mu) &= l(\lambda)^T J\, Bl(\mu) = \lambda^{n-1} l(\lambda^{-1})^T\, Bl(\mu) \\ &= \frac{1}{\lambda - \mu} b(\lambda^{-1}, \mu)\lambda^n. \qquad \square \end{aligned}$$

THEOREM 4.1. *Suppose that A is an r-Bezoutian,*

$$(\lambda - \mu)A(\lambda, \mu) = \sum_{i=1}^{r} u_i(\lambda)v_i(\mu),$$

and α, β fixed numbers with $\alpha\beta \neq 1$. Then

$$A = \frac{1}{1 - \alpha\beta} \sum_{i=1}^{r} JC_\alpha(\hat{u}_i)^T C_\beta(v_i), \tag{4.5}$$

where $\hat{u}_i(\lambda) := u_i(\lambda^{-1})\lambda^n$.

Proof. Let A' denote the right-hand side of (4.5). Then $B := (1 - \alpha\beta)JA'$ is of the form (4.2), where $g_i = \hat{u}_i$, $h_i = v_i$. Since $\sum u_i(\lambda)v_i(\lambda) = 0$ for all λ, the condition (4.3) is fulfilled. Thus, by Lemma 4.1,

$$B(\lambda, \mu) = \frac{1 - \alpha\beta}{1 - \lambda\mu} \sum_{i=1}^{r} \hat{u}_i(\lambda)v_i(\mu).$$

It remains to apply Lemma 4.2 in order to get (4.5). □

Let us note that, for $\alpha \neq 0$,

$$C_\alpha(\hat{u})^T = \alpha C_{1/\alpha}(u).$$

Taking this relation into account, we obtain from (4.5) for $\beta \neq 0$, letting $\alpha \to \infty$, the formula

$$A = -\frac{1}{\beta} \sum_{i=1}^{r} JC_0(u_i)C_\beta(v_i). \tag{4.6}$$

Furthermore, we get for $\alpha \neq 0$, letting $\beta \to \infty$, the formula

$$A = -\frac{1}{\alpha} \sum_{i=1}^{r} JC_\alpha(\hat{u}_i)^T C_0(\hat{v}_i)^T. \tag{4.7}$$

Letting both $\alpha \to \infty$ and $\beta \to \infty$, we obtain

$$A = -\sum_{i=1}^{r} JC_0(u_i)C_0(\hat{v}_i)^T. \tag{4.8}$$

COROLLARY 4.1. *For A given in Theorem* 4.1, *the formulas* (4.6), (4.7), *and* (4.8) *hold.*

Let us note that in formula (4.8) only triangular Toeplitz and Hankel matrices are involved. Therefore, it is, in a sense, a generalization of the well-known Gohberg–Semencul formula.

5. Characterization of the parameters. In § 3 we have proved that the MPI H^+ of a Hankel matrix H has a generating function of the form

$$H^+(\lambda, \mu) = \frac{1}{\lambda - \mu} \sum_{i=1}^{4} u_i(\lambda) v_i(\mu). \tag{5.1}$$

In § 4 we have shown that this formula leads to matrix representations, which can be utilized for the fast computation of pseudo-solutions. In this section, we discuss the problem of how the parameters u_i and v_i can be characterized. We present several possibilities.

For simplicity of presentation, we restrict ourselves to the case of an $m \times n$ Hankel matrix with rank $H < \min\{m, n\}$. This includes the singular square case. The full rank case can be considered in a similar way.

5.1. First we recall how the vectors u_i and v_i ($i = 1, 2, 3, 4$) appear in the proof of Theorem 1.1 in § 3. According to (3.3), the relation (5.1) holds with u_i, v_i being

$$u_1 = x_1, \quad u_2 = x_2, \quad u_3(\lambda) = w_1(\lambda) - \lambda^n, \quad u_4 = w_2,$$
$$v_1(\lambda) = \tilde{w}_1(\lambda) - \lambda^m, \quad v_2 = \tilde{w}_2, \quad v_3 = -\tilde{x}_1, \quad v_4 = -\tilde{x}_2.$$

Here x_i, w_i, $\tilde{x}_i$, and $\tilde{w}_i$ denote the first parts, i.e., the first n components, of the solutions of the equations

$$\begin{gathered} \mathscr{H}\begin{bmatrix} x_1 \\ \xi_1 \end{bmatrix} = \begin{bmatrix} e(m) \\ 0 \end{bmatrix}, \qquad \mathscr{H}\begin{bmatrix} x_2 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 0 \\ e(\mathscr{H}) \end{bmatrix}, \\ \mathscr{H}\begin{bmatrix} w_1 \\ \eta_1 \end{bmatrix} = \begin{bmatrix} g(m, H) \\ 0 \end{bmatrix}, \qquad \mathscr{H}\begin{bmatrix} w_2 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} g(m, M_m(\bar{u})) \\ 0 \end{bmatrix}, \end{gathered} \tag{5.2}$$

and the analogous equations for $\mathscr{H}^T$.

In order to have regard for the corresponding equations for $\mathscr{H}$ and $\mathscr{H}^T$, we introduce the notation $\tilde{u}_1 = -v_3$, $\tilde{u}_2 = -v_4$, $\tilde{u}_3 = v_1$, $\tilde{u}_4 = v_2$. We introduce furthermore the polynomial matrices

$$U(\lambda) = [u_1(\lambda) \quad u_2(\lambda) \quad u_3(\lambda) \quad u_4(\lambda)],$$
$$\tilde{U}(\lambda) = [\tilde{u}_1(\lambda) \quad \tilde{u}_2(\lambda) \quad \tilde{u}_3(\lambda) \quad \tilde{u}_4(\lambda)],$$

and the signature matrix

$$\Theta = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}.$$

Then the relation (5.1) can be written in a compact form, i.e.,

$$H^+(\lambda, \mu) = \frac{1}{\lambda - \mu} U(\lambda)\Theta\tilde{U}(\mu)^T. \tag{5.3}$$

5.2. We characterize now the vectors occurring in (5.1) in terms of pseudo-solutions for the matrix H. In view of Theorem 2.2, we have

$$x_1 = H^+e(m), \quad w_1 = H^+g(m, H), \quad w_2 = H^+g(m, M_m(\bar{u})),$$

and

$$x_2 = (M_n(\bar{u})^T)^+ e(\varkappa).$$

We introduce the matrix $T_\varkappa = M_n(\bar{u})^T M_n(u)$, which is a positive definite Toeplitz matrix. Then we have

$$(M_n(\bar{u})^T)^+ = M_n(u) T_\varkappa^{-1}.$$

Hence

$$x_2 = M_n(u) T_\varkappa^{-1} e(\varkappa).$$

That means H^+ can be constructed with the help of three pseudo-solutions and the last column of a Toeplitz matrix inverse.

5.3. A more general characterization will be obtained using the solutions of certain homogeneous equations. To explain this we need some notation.

For an $r \times s$ Hankel matrix $H_0 = [h_{k+l}]_0^{r-1}{}_0^{s-1}$, ∂H_0 will denote the $(r-1) \times (s+1)$ Hankel matrix $[h_{k+l}]_0^{r-2}{}_0^{s}$. The matrix ∂H_0 is one member of the family of Hankel matrices generated by H_0 as it is described in § 2. If $\mathscr{H} = [H_{ij}]$ is a Hankel mosaic matrix, then we define $\partial\mathscr{H} = [\partial H_{ij}]$. In particular, for $\mathscr{H}$ defined by (2.3), $\partial\mathscr{H}$ will be a full rank matrix with kernel dimension 4.

Observation. The vectors

$$U_1 = \begin{bmatrix} x_1 \\ 0 \\ \xi_1 \\ 0 \end{bmatrix}, \quad U_2 = \begin{bmatrix} x_2 \\ 0 \\ \xi_2 \\ 0 \end{bmatrix}, \quad U_3 = \begin{bmatrix} w_1 \\ -1 \\ \eta_1 \\ 0 \end{bmatrix}, \quad U_4 = \begin{bmatrix} w_2 \\ 0 \\ \eta_2 \\ -1 \end{bmatrix} \tag{5.4}$$

form a basis of ker $\partial\mathscr{H}$.

Now let $Z_i = [z_i^T \zeta_i^T]^T$ $(i = 1, 2, 3, 4)$ be an arbitrary basis of ker $\partial\mathscr{H}$. Then there is a regular 4×4 matrix C such that $[U_1\ U_2\ U_3\ U_4] = [Z_1\ Z_2\ Z_3\ Z_4]C$.

The matrix C can be determined as follows. Let Φ denote the matrix with the four rows f_i $(i = 1, 2, 3, 4)$, where

$$f_1 = [s_{m-1} \cdots s_{m+n-1}\ \bar{u}_p\ 0 \cdots 0],$$

$$f_2 = [\bar{u}_0 \cdots \bar{u}_p\ 0 \cdots 0\ 0 \cdots 0],$$

$f_3 = -[e(n+1)^T\ 0]^T$, and $f_4 = -[0\ e(\mathscr{H}+1)^T]^T$. Then $\Phi[U_1\ U_2\ U_3\ U_4] = I_4$ and $P := \Phi[Z_1\ Z_2\ Z_3\ Z_4]$ is nonsingular. Now we have $P^{-1} = C$.

Denoting $Z(\lambda) = [z_1(\lambda)\ z_2(\lambda)\ z_3(\lambda)\ z_4(\lambda)]$, (5.3) goes over into

$$H^+(\lambda, \mu) = \frac{1}{\lambda - \mu} Z(\lambda) P^{-1} \Theta \tilde{P}^{-1} \tilde{Z}(\mu)^T,$$

where $\tilde{Z}(\mu)$ and $\tilde{P}$ are analogously defined as $Z(\lambda)$ and P. So we have proved that H^+ is determined by any basis of ker $\partial\mathscr{H}$ and ker $\partial\mathscr{H}^T$.

Remark. It can be checked that already ker $\partial\mathscr{H}$ determines H^+.

5.4. We show now that two vectors belonging to the kernel of $\partial\mathscr{H}$ can easily be determined. For these vectors, the last components vanish. Suppose that $[x\ 0]^T \in$ ker $\partial\mathscr{H}$, $x \in \mathbb{C}^{n+1}$. Then $x \in$ ker ∂H and $x \in$ ker $M_{n+1}(\bar{u}')^T$, where $u' = [0\ u^T\ \ 0]^T$, and vice versa. According to Theorem 2.1, we have $x \in$ ker ∂H if and only if there is a

$g \in \mathbb{C}^{\varkappa+1}$ such that $x = M_{n+1}(u)g$. Hence $[x\ 0]^T$ belongs to ker $\partial\mathscr{H}$ if and only if x is of this form and $T'g = 0$, where T' is the $(\varkappa - 1) \times (\varkappa + 1)$ Toeplitz matrix

$$T' = M_{n+1}(\bar{u}')^T M_{n+1}(u).$$

Let us note that $T_\varkappa$ defined above goes over into T', applying a transformation which is analogous to the transformation ∂ for Hankel matrices. The kernel of T' is two-dimensional and can be described with the help of the first column of $T_\varkappa^{-1}$. In fact, let c be this column. Then

$$g_1 = \begin{bmatrix} c \\ 0 \end{bmatrix}, \qquad g_2 = \begin{bmatrix} 0 \\ \hat{\bar{c}} \end{bmatrix}$$

is a basis of ker T'. Recall that $\hat{c}$ denotes the reflection of c. Hence

$$V_i = \begin{bmatrix} M_{n+1}(u)g_i \\ 0 \end{bmatrix} \qquad (i = 1, 2)$$

belong to ker $\partial\mathscr{H}$. Furthermore, these two vectors are linearly independent.

In order to get a basis of ker $\partial\mathscr{H}$, one has to add two of the vectors U_i defined by (5.4). The question we still have to answer is that of linear independence. U_1 is obviously linearly independent of the V_i if $\xi_1 \neq 0$. According to (2.4),

$$\xi_1 = M_m(\bar{u})^+ e(m) = \bar{T}_{-\nu}^{-1} M_m(u)^T e(m).$$

From this relation we conclude that $\xi_1 \neq 0$ if and only if the last component of u is nonzero. So we have proved that H^+ is determined by the vectors U_1, U_4, and c if $u_p \neq 0$. In [13] singular matrices with this property are called quasi-regular.

It can be shown that when $u_p = 0$, the vector U_1 can be replaced by U_2 in order to get a basis of ker $\partial\mathscr{H}$.

5.5. Finally, we discuss the problem of whether H^+ can be constructed with the help of some columns and rows of H^+. First we observe that the vector x_1 occurring in (5.1) and (5.3) equals the last column of H^+. Let x_0 denote the first column of H^+ and let ξ_0 be such that

$$\mathscr{H} \begin{bmatrix} x_0 \\ \xi_0 \end{bmatrix} = \begin{bmatrix} e_0 \\ 0 \end{bmatrix},$$

where e_0 denotes the first unit vector in $\mathbb{C}^m$. Then $[0\ x_0^T\ 0\ \xi_0^T]^T$ belongs to ker $\partial\mathscr{H}$. We have

$$\xi_0 = M_m(\bar{u})^+ e_0 = \bar{T}_{-\nu}^{-1} M_m(u)^T e_0.$$

Hence $\xi_0 \neq 0$ if $u_0 \neq 0$. Thus we obtained that the vectors ξ_0 and ξ_1 and, therefore also the corresponding elements of ker $\partial\mathscr{H}$, are linearly independent if $u_0 \neq 0$ and $u_p \neq 0$. So we got the following result.

THEOREM 5.1. *Let H be an $m \times n$ Hankel matrix with* rank $H < \min\{m, n\}$ *and let u be a vector generating the kernel of H as described above. If the first and the last components of u are nonzero, then the matrix H^+ can be constructed from its first and last columns and rows and the first column of the inverse of the Toeplitz matrix $T_\varkappa$.*

6. Proof of Theorem 1.3 and MPIs of circulants. In this section we restrict ourselves to square Hankel matrices. Our main aim is the proof of Theorem 1.3. Since, as noted above, the inverse of a nonsingular Hankel matrix is a 2-Bezoutian, we may assume that

H is singular. We split the proof of Theorem 1.3 into the necessity and sufficiency parts. First, we shall prove the following.

THEOREM 6.1. *If H is a singular square Hankel matrix such that the* MPI H^+ *has Bezoutian rank less than* 4, *then H is a Hankel circulant matrix.*

Proof. According to (3.3), we have

$$\nabla H^+ = \sum_{i=1}^{4} u_i v_i^T, \tag{6.1}$$

where

$$u_1 = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}, \quad u_2 = \begin{bmatrix} x_2 \\ 0 \end{bmatrix}, \quad u_3 = \begin{bmatrix} w_1 \\ -1 \end{bmatrix}, \quad u_4 = \begin{bmatrix} w_2 \\ 0 \end{bmatrix},$$

$$v_1 = u_3, \quad v_2 = u_4, \quad v_3 = -u_1, \quad v_4 = -u_2,$$

and x_i, w_i denote the first parts of the solutions of (5.2).

According to the assumptions, we have rank $\nabla H^+ < 4$. That means the vectors u_1, $\ldots$, u_4 are linearly dependent. Since the last components of u_1, u_2, u_4 vanish and the last component of u_3 equals -1, the vectors u_1, u_2, u_4 are already linearly dependent. Hence there is a nontrivial vanishing linear combination, i.e.,

$$\alpha x_1 + \beta x_2 - \gamma w_2 = 0.$$

We now apply the operator $M^T := M_n(\bar{u})^T$ to this equality. In view of (2.3) and (5.2), we have

$$M^T x_1 = 0, \quad M^T x_2 = e_{\varkappa-1}, \quad M^T w_2 = 0.$$

Hence $\beta e_{\varkappa-1} = 0$, i.e., $\beta = 0$. So we have proved that the vectors x_1 and w_2 are linearly dependent.

According to the definition of x_1 and w_2, we have $x_1 = H^+ e_{n-1}$, $w_2 = H^+ S\bar{u}$, where S denotes the backward shift operator, i.e.,

$$(S\bar{u})(\lambda) = \frac{\bar{u}(\lambda) - \bar{u}(0)}{\lambda}.$$

In view of $\alpha x_1 - \gamma w_2 = 0$, we have $\alpha e_{n-1} - \gamma S\bar{u} \in \ker H^+$. But $\ker H^+ = \ker H^* = \ker \bar{H}$. Hence $\alpha\lambda^{n-1} - \gamma(S\bar{u})(\lambda) \in \mathscr{C}_n$. Now we apply the kernel structure Theorem 2.1 and conclude that there is a polynomial $q(\lambda)$ such that

$$\alpha\lambda^{n-1} - \gamma\frac{\bar{u}(\lambda) - \bar{u}(0)}{\lambda} = q(\lambda)\bar{u}(\lambda).$$

From this relation we see that $\alpha \neq 0$ or $u(0) \neq 0$; otherwise we would have a contradiction to Theorem 2.1. Furthermore, we obtain

$$\alpha\lambda^n - \gamma\bar{u}(0) = (\gamma + \lambda q(\lambda))\bar{u}(\lambda).$$

By Theorem 2.1, this implies $\alpha\lambda^n - \gamma\bar{u}(0) \in \mathscr{C}_{n+1}$. In other words, $\alpha e_n - \gamma\bar{u}(0)e_0$ belongs to the kernel of $\bar{H}_{n+1}$. That means we have

$$\alpha\bar{s}_{i+n} = \gamma\bar{u}(0)\bar{s}_i \qquad (i = 0, \ldots, n-2)$$

or, equivalently,

$$\bar{\alpha}s_{i+n} = \bar{\gamma}u(0)s_i.$$

Taking into account that one of the numbers α or γ and one of the numbers α or $u(0)$ are different from zero, one concludes that H is a t-Hankel circulant for $t = \bar{\gamma}u(0)/\bar{\alpha}$, $t = 0$, or $t = \infty$. □

We still have to prove the sufficiency part of Theorem 1.3, which means we have to investigate the MPI of Hankel circulant matrices.

Let us start with an observation concerning Hankel circulants. Any α-Hankel circulant H, $\alpha \neq \infty$, can be represented in the form

$$H = H_0(g) + \alpha H_\infty(g), \tag{6.2}$$

where

$$H_0(g) = \begin{bmatrix} g_1 & \cdots & g_n \\ \vdots & \cdot^{\cdot^{\cdot}} & \\ g_n & & 0 \end{bmatrix}, \qquad H_\infty(g) = \begin{bmatrix} 0 & & g_0 \\ & \cdot^{\cdot^{\cdot}} & \vdots \\ g_0 & \cdots & g_{n-1} \end{bmatrix}.$$

In the case $\alpha = \infty$, we have $H = H_\infty(g)$. It is easily checked that

$$H_0(g) = \operatorname{Bez}(g, 1), \qquad H_\infty(g) = -\operatorname{Bez}(g, \lambda^n),$$

where $g(\lambda) = g_0 + g_1\lambda + \cdots + g_n\lambda^n$. Hence the following is true.

PROPOSITION 6.1. *An α-Hankel circulant matrix H can be represented in the form*

$$H = \operatorname{Bez}(g, 1 - \alpha\lambda^n) \quad \textit{if } \alpha \neq \infty \quad \textit{or} \quad H = \operatorname{Bez}(g, -\lambda^n) \quad \textit{if } \alpha = \infty.$$

Now we are going to prove the following theorem, which will be formulated for general Hankel circulants.

THEOREM 6.2. *The* MPI *of a Hankel circulant matrix H is a classical Bezoutian.*

Proof. For the case of a nonsingular matrix H, the assertion follows from the well-known facts that the inverse of a circulant is a circulant again and that a Hankel circulant is a classical Bezoutian (cf. § 1).

Now let H be a singular α-Hankel circulant. First, we consider the case $\alpha \neq \infty$. Obviously, the vector $e_n - \alpha e_0$ $(\in \mathbb{C}^{n+1})$ belongs to $\ker \partial H$ (∂H was introduced in § 5.3) or, what is the same, $\lambda^n - \bar{\alpha} \in \mathscr{C}_{n+1}$. Let $u(\lambda)$ denote the polynomial occurring in Theorem 2.1. Then, according to this theorem,

$$\lambda^n - \bar{\alpha} = q(\lambda)\bar{u}(\lambda)$$

for a certain polynomial $q(\lambda)$. Hence $\bar{\alpha} = -q(0)\bar{u}(0)$ and

$$\lambda^n = q(\lambda)\bar{u}(\lambda) - q(0)\bar{u}(0),$$

which implies

$$\lambda^{n-1} - q(0)\frac{\bar{u}(\lambda) - \bar{u}(0)}{\lambda} = q_1(\lambda)\bar{u}(\lambda),$$

where

$$q_1(\lambda) = \frac{q(\lambda) - q(0)}{\lambda}.$$

Thus $e_{n-1} - q(0)S\bar{u}$ belongs to $\ker \bar{H} = \ker H^+$.

Now we take into account that $H^+e_{n-1} = x_1$ and $H^+S\bar{u} = w_2$, where x_1 and w_2 are involved in the representation (6.1). We obtain $x_1 = q(0)w_2$, which implies

$$u_1 = q(0)u_4, \qquad v_3 = -q(0)v_2.$$

Now (6.1) goes over into

$$\begin{aligned}\nabla H^+ &= u_4(q(0)v_1 + v_4)^T + (u_2 - q(0)u_3)v_2^T \\ &= u_4(q(0)u_3 - u_2)^T - (-u_2 + q(0)u_3)u_4^T.\end{aligned}$$

Hence H^+ is the Bezoutian of $u_4(\lambda)$ and $q(0)u_3(\lambda) - u_2(\lambda)$.

It remains to consider the case $\alpha = \infty$. In this situation, the function that is identically equal to 1 belongs to $\mathscr{C}_{n+1}$ or, what is the same,

$$1 = \bar{q}(\lambda)\bar{u}(\lambda)$$

for a polynomial $\bar{q}(\lambda)$. Hence $u(\lambda)$ is a constant, which implies $w_2 = 0$, due to (5.2), and $u_4 = v_2 = 0$. In this case, (6.1) looks as follows:

$$\nabla H^+ = u_1 v_1^T - v_1 u_1^T = u_1 u_3^T - u_3 u_1^T.$$

This shows that H^+ is a classical Bezoutian. □

Acknowledgment. The authors thank an anonymous referee for helpful comments and criticism.

REFERENCES

[1] G. Ammar and P. Gader, *New decompositions of inverses of a Toeplitz matrix*, in Signal Processing, Scattering and Operator Theory, and Numerical Methods, Proc. of the International Symposium MTNS-89, Vol. 3, Birkhäuser, Boston, 1990.

[2] B. Anderson and E. Jury, *Generalized Bezoutian and Sylvester matrices in multivariable linear control*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 551–556.

[3] A. Ben-Artzi and T. Shalom, *On inversion of block Toeplitz matrices*, Integral Equations Operator Theory, 8 (1985), pp. 173–192.

[4] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, John Wiley, New York, 1974.

[5] P. J. Davis, *Circulant Matrices*, John Wiley, New York, 1979.

[6] P. Delsarte, Y. Genin, and Y. Kamp, *On the Toeplitz embedding of an arbitrary matrix*, Linear Algebra Appl., 51 (1983), pp. 97–119.

[7] I. Gohberg and G. Heinig, *Inversion of finite-section Toeplitz matrices consisting of elements of a noncommutative algebra*, Rev. Roumaine Math. Pures Appl., 19 (1974), pp. 623–633. (In Russian.)

[8] I. Gohberg and A. A. Semencul, *On the inversion of finite-section Toeplitz matrices and their continuous analogues*, Mat. Issled., 7 (1972), pp. 201–224. (In Russian.)

[9] I. Gohberg and T. Shalom, *On the Bezoutian of non-square matrix polynomials and inversion of matrices with non-square blocks*, Linear Algebra Appl., 137/138 (1990), pp. 249–324.

[10] T. N. E. Greville, *Toeplitz matrices with Toeplitz inverses*, Linear Algebra Appl., 55 (1983), pp. 87–92.

[11] G. Heinig, *Matrix representations of Bezoutians*, Linear Algebra Appl., submitted.

[12] G. Heinig and P. Jankowski, *Parallel and superfast algorithms for Hankel systems of equations*, Numer. Math., 58 (1990), pp. 109–127.

[13] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-Like Matrices and Operators*, Akademie-Verlag, Berlin, and Birkhäuser-Verlag, Basel, Boston, Stuttgart, 1984.

[14] G. Heinig and A. Tewodros, *On the inverses of Hankel and Toeplitz mosaic matrices*, in Seminar Analysis: Operator Equations and Numerical Analysis, Karl-Weierstraß-Institut, Berlin, 1988, pp. 53–65.

[15] F. Hellinger, *Moore-Penrose-Invertierung von Hankelmatrizen*, Diplomarbeit, Universität Leipzig, Leipzig, Germany, 1990.

[16] N. M. Hung and R. E. Cline, *Inversion of persymmetric matrices having Toeplitz inverses*, J. Assoc. Comput. Mach., 19 (1972), pp. 437–444.

[17] T. Kailath, S. Y. Kung, and M. Morf, *Displacement ranks of a matrix*, Bull. Amer. Math. Soc., 1 (1979), pp. 769–773.

[18] T. Kailath and J. Chun, *Generalized Gohberg–Semencul formulas for matrix inversion*, in Operator Theory: Advances and Applications, Vol. 40, Birkhäuser-Verlag, Basel, 1989.

[19] M. G. Krein and M. A. Naimark, *The method of symmetric and Hermitian forms in the theory of separation of the roots of algebraic equations*, Linear and Multilinear Algebra, 10 (1981), pp. 265–308.

[20] F. I. Lander, *The Bezoutian and the inversion of Hankel and Toeplitz matrices*, Mat. Issled., 9 (1974), pp. 69–87. (In Russian.)

[21] L. Lerer and M. Tismenetsky, *Generalized Bezoutians and the inversion problem for block Toeplitz matrices*, Integral Equations Operator Theory, 9 (1986), pp. 790–819.

[22] ———, *Generalized Bezoutians and matrix equations*, Linear Algebra Appl., 96 (1988), pp. 123–160.

[23] T. Shalom, *On algebras of Toeplitz matrices*, Linear Algebra Appl., 128 (1987), pp. 211–226.

[24] H. K. Wimmer, *Generalized Bezoutians and block Hankel matrices*, Linear Algebra Appl., 117 (1989), pp. 105–113.

[25] ———, *On the history of the Bezoutian and resultant matrix*, Linear Algebra Appl., 128 (1990), pp. 27–34.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 646–654, July 1993

003

DISPLACEMENT RANK OF GENERALIZED INVERSES OF PERSYMMETRIC MATRICES*

PIERRE COMON† AND PASCAL LAURENT-GENGOUX‡

Abstract. Toeplitz matrices are persymmetric matrices belonging to the large class of so-called structured matrices, characterized by their displacement rank. This characterization was introduced 12 years ago by Kailath and others. In this framework, properties of singular structured persymmetric matrices are investigated with the goal of proving the possible existence of fast algorithms for computing their pseudo-inverses. Loosely speaking, it is proved that the pseudo-inverses of some structured matrices with displacement rank r have a displacement rank bounded by $2r$.

Key words. Toeplitz, structure, persymmetric, singular, Schur algorithm, pseudo-inverse

AMS subject classifications. 15A09, 15A57, 65F30

1. Introduction.

Toeplitz matrices and more generally structured matrices are encountered in several problems, including prediction of almost stationary processes, modeling of stochastic processes by state-space systems, lossless transmission lines, localization in antenna array processing, or testing relative primeness of polynomials [1], [2], [5], [6]. Because of their sometimes very large size, the structured linear systems must be solved by resorting to specialized algorithms taking advantage of their particular features in order to reduce both computational load and storage requirements, without disregarding the possibilities of parallel implementations.

There exist numerous fast algorithms for solving linear Toeplitz systems, and the most well known are the Levinson and Schur algorithms. Extensions have been proposed since 1979 for matrices whose structure was close to Toeplitz in some sense [2], [3]. For full-rank matrices, the main results may be found in [3]–[5] and [9]. Nevertheless, the leading minors are required to be nonzero in order for existing fast algorithms to be stable [2]. Improvements have been proposed in order to allow fast algorithms to run in a stable way for the larger class of regular matrices with arbitrary rank profile [12], [14]. Besides this first limitation, it is now proved that the proximity of a regular matrix to the Toeplitz structure is preserved by inversion, but nothing is known to date regarding singular matrices. Singular structured matrices will be the subject of our discussion, and we shall focus our attention on persymmetric matrices (relevant in the Toeplitz case, for instance). It will be proved that the generalized inverse of a singular structured persymmetric matrix has a structure that could be defined with the help of generators, exactly as in the regular case [3]; in other words, its displacement rank is bounded. Our approach, however, is not constructive in the sense that no means is provided to obtain explicitly the corresponding generators. Only their existence is proved. This result is important since it demonstrates that the generalized inverse of an $N \times N$ close to Toeplitz matrix can be completely described only by a restricted number of vectors of size N (four in the case of a singular Toeplitz matrix).

The paper is organized as follows. Section 2 defines notations and states mostly known results for regular matrices. The body of the paper is § 3, in which properties of some singular symmetric persymmetric structured matrices are investigated. Section 4

* Received by the editors November 19, 1990; accepted for publication (in revised form) December 3, 1991.

† Thomson-Sintra, BP 157, F-06903 Sophia-Antipolis Cedex, France (na.comon@na-net.ornl.gov).

‡ Ecole Centrale Paris, 92295 Chatenay Malabry Cedex, France. Present address, Thomson-Sintra, BP 157, F-06903 Sophia-Antipolis Cedex, France.

briefly extends the result of § 3 to the nonsymmetric case. Although Theorem 3.1 is a particular case of Theorem 4.1 (§ 3 could be partially skipped), we found it clearer to go first through the less general symmetric case.

2. Displacement of matrices. The concept of matrix displacement has been introduced in [2], and discussed in a more general framework in [5]. Only a few definitions and notations are recalled below for convenience, and readers are invited to consult the above-mentioned references as well as [1], [3], and [4] for more details.

DEFINITION 2.1. Let Z be a fixed nilpotent square matrix. Two types of displaced matrices will be used; for any square matrix T, they are defined as

$$\nabla T = T - ZTZ \quad \text{and} \quad \Delta T = T - Z^t T Z. \tag{2.1}$$

The rank of matrix ∇T (respectively, ΔT) is called the displacement rank of T with respect to Z (respectively, Z^t).

DEFINITION 2.2. Any set of pairs of vectors $\{(x_1, y_1), (x_2, y_2), \ldots, (x_q, y_q)\}$, satisfying

$$\nabla T = \sum_{i=1}^{q} \alpha_i x_i y_i^t, \qquad \alpha_i \in \{-1, 1\}, \tag{2.2}$$

is a set of generators. If the matrix T is symmetric, then we can take $x_i = y_i$ and the sequence of signs $\{\alpha_i\}$ is called the displacement signature of T.

If q is minimum, i.e., if q is equal to the displacement rank of T, then those generators are called minimal [7]. Only minimal generators will be referred to in the rest of the paper. The Crout decomposition of ∇T is one way for building minimal generators.

Property 2.1. As indicated by their name, generators can be used to recover the original matrix. In fact, it suffices to form the sum

$$T = \sum_{i=0}^{N-1} Z^i \nabla T Z^{it}. \tag{2.3}$$

Another equivalent expression may be obtained by stacking the successive shifted generators in triangular matrices,

$$T = \sum_{i=1}^{q} \alpha_i L_i U_i, \quad L_i = [x_i \, Zx_i \cdots Z^{N-1} x_i], \quad U_i = [y_i \, Zy_i \cdots Z^{N-1} y_i]^t. \tag{2.4}$$

This property shows that if a matrix T has a small displacement rank compared to its size, then it may be stored efficiently under the form of $2q$ generators (q are sufficient if T is symmetric).

DEFINITION 2.3. For the sake of simplicity, the displacements of the inverse matrix T^{-1} will be denoted in short as ∇T^{-1} and ΔT^{-1}, standing for $\nabla(T^{-1})$ and $\Delta(T^{-1})$. Inverses of displaced matrices are not used in the paper, thus avoiding any confusion.

THEOREM 2.1. *If T is a nonsingular square symmetric matrix, displaced matrices ∇T and ΔT^{-1} have the same rank. The same property holds for ΔT and ∇T^{-1}.*

This result is attributed to Gohberg and Semencul and can be traced back to 1972. The first proof given below can be found in [1], [4], and [5], and has some interest because of its conciseness. Useful relations with orthogonal polynomials are also stressed in [10], [11]. We derive then a second proof, more convenient for further extensions to the case of singular matrices. This emphasizes the differences between the principles utilized.

First proof. The shortest proof that can be given consists of writing the matrix

$$\begin{pmatrix} T & Z \\ Z^t & T^{-1} \end{pmatrix}$$

in two different manners:

$$\begin{pmatrix} I & ZT \\ 0 & I \end{pmatrix}\begin{pmatrix} \nabla T & 0 \\ 0 & T^{-1} \end{pmatrix}\begin{pmatrix} I & 0 \\ TZ^t & I \end{pmatrix} \text{ and } \begin{pmatrix} I & 0 \\ Z^tT^{-1} & I \end{pmatrix}\begin{pmatrix} T & 0 \\ 0 & \Delta T^{-1} \end{pmatrix}\begin{pmatrix} I & T^{-1}Z \\ 0 & I \end{pmatrix}.$$

Yet, from the Sylvester lemma, inertia is preserved by congruent transformations, which shows that ΔT^{-1} and ∇T have the same inertia, since so do T^{-1} and T. Note that this result is stronger than the theorem.

Second proof. We shall prove in three steps that the null spaces of ΔT^{-1} and ∇T have the same dimension. Denote $U = \text{Ker } \nabla T$ and $V = TZ^t(U)$ the image of U by the operator TZ^t. We prove first that

$$V \subset \text{Ker } \Delta T^{-1}. \tag{2.5}$$

Let $u \in U$ and $v = TZ^tu$. Then

$$\Delta T^{-1}v = T^{-1}TZ^tu - Z^tT^{-1}ZTZ^tu.$$

But since $u \in \text{Ker } \nabla T$,

$$ZTZ^tu = Tu,$$

and hence

$$\Delta T^{-1}v = Z^tu - Z^tT^{-1}Tu = 0,$$

which proves (2.5). Next TZ^t is one-to-one on U. In fact, if $TZ^tu = 0$ for $u \in U$, then $Tu = ZTZ^tu = 0$, implying $u = 0$ because T is nonsingular.

This implies that $\dim U = \dim V$, which together with (2.5) and the definition of U yields

$$\dim \nabla T \leqq \dim \text{Ker } \Delta T^{-1}.$$

A similar argument with T and Z^t replaced by T^{-1} and Z, and with ∇ and Δ interchanged, implies the reverse inclusion. This completes the proof. □

THEOREM 2.2. *Let the matrix Z denote the so-called "lower shift" matrix*

$$Z = \left(\begin{array}{cc:c} & 0 & 0 \\ \cdot\cdot & \cdot\cdot & \cdot\cdot \\ & I_{N-1} & 0 \\ & & \end{array}\right).$$

Then a symmetric Toeplitz matrix T always has a displacement rank 2 with respect to Z. Moreover, if diagonal entries of T are normalized to 1, the range of ∇T is spanned by t_1 and ZZ^tt_1, where t_1 denotes the first column of T. The vectors t_1 and ZZ^tt_1 are minimal generators of ∇T.

Proof. See [4] or [7] for a proof.

3. Displacement rank of Moore–Penrose inverses.

DEFINITION 3.1. Define the generalized inverse of any square matrix T, denoted T^-, as the matrix satisfying the four Moore–Penrose conditions: (i) $TT^-T = T$, (ii) $T^-TT^- = T^-$, (iii) $(TT^-)^t = TT^-$, and (iv) $(T^-T)^t = T^-T$.

The null space of the generalized inverse T^- is the null space of T^t and the range of T^- is the range of T^t [8]. Hence in the symmetric case the generalized inverse T^- has the same range and null spaces as T.

For the sake of simplicity, we shall only concern ourselves with the symmetric case in this section. The nonsymmetric case will be postponed to § 4.

DEFINITION 3.2. Define the backward identity matrix J as

$$J_{i,j} = 0 \quad \text{except } J_{i,N-i+1} = 1.$$

J is sometimes called the anti-identity matrix, or the reverse unit matrix. A matrix M will be called persymmetric if it satisfies

$$JMJ = M^t.$$

From now on the displacement matrix Z will be assumed to be persymmetric.

THEOREM 3.1. *Let T be an $N \times N$ symmetric and persymmetric matrix. Then if r is the displacement rank of T, the displacement rank of its generalized inverse T^-, is bounded by $2r$.*

Let us examine step by step the proof derived above in the nonsingular case for T^{-1}. The first obstacle in extending it to the generalized inverse of T is that we cannot use the property $T^{-1}T = I_N$, but only $TT^-T = T$; this means that the main step of the proof fails, namely, that $TZ^t(\text{Ker } \nabla T) \subset \text{Ker } \Delta T^{-1}$.

Nevertheless, we can still prove, as we shall see, that the quadratic form associated with ΔT^- (and not the linear operator itself) vanishes on a subspace E of dimension $N - r$; from this it follows that the dimension of Ker ΔT^- is smaller than $N - 2r$, as will be shown subsequently. This subspace E is built as a sum of two subspaces, $V \oplus W$, where $V = TZ^t(\text{Ker } \nabla T)$ and W is a subspace of Ker T. The proof of Theorem 3.1 requires two lemmas.

LEMMA 3.1. *The quadratic form $\langle \Delta T^- v, v\rangle$ vanishes for all $v \in V$, $V = TZ^t(\text{Ker } \nabla T)$.*

Proof. Let u be in Ker ∇T and $v = TZ^t u$. Let us calculate $\langle \Delta T^- v, v\rangle$:

$$\langle \Delta T^- v, v\rangle = \langle T^- TZ^t u, TZ^t u\rangle - \langle Z^t T^- ZTZ^t u, TZ^t u\rangle.$$

By using the transpose rule for operators and the symmetry of T, we get

$$\langle \Delta T^- v, v\rangle = \langle TT^- TZ^t u, Z^t u\rangle - \langle T^- ZTZ^t u, ZTZ^t u\rangle.$$

But $ZTZ^t u = Tu$ for $u \in \text{Ker } \nabla T$, and $TT^-T = T$ by definition of T^-. Hence

$$\langle \Delta T^- v, v\rangle = \langle TZ^t u, Z^t u\rangle - \langle T^- Tu, Tu\rangle,$$

and resorting to transposition gives

$$\langle \Delta T^- v, v\rangle = \langle ZTZ^t u, u\rangle - \langle TT^- Tu, u\rangle.$$

Now using the properties $ZTZ^t u = Tu$ and $TT^-T = T$ again yields

$$\langle \Delta T^- v, v\rangle = \langle Tu, u\rangle - \langle Tu, u\rangle = 0. \qquad \square$$

LEMMA 3.2. *Define K as the null space of the operator TZ^t in* Ker ∇T, *i.e.*,

$$K = \text{Ker } TZ^t \cap \text{Ker } \nabla T.$$

Define also the subspace

$$W = J(K).$$

With the notations defined above, we have

(i) $W \subset \operatorname{Ker} \Delta T^-$,

(ii) $\dim (W) = \dim (K)$, *and*

(iii) $W \subset \operatorname{Ker} T$.

Proof. Symmetry and persymmetry imply centrosymmetry:

$$T = JTJ,$$

while the following identities are also satisfied:

$$Z = JZ^tJ \quad \text{and} \quad J^2 = I_N.$$

First, let us check that W is included in $\operatorname{Ker} \Delta T^-$. Let x be in K and note that this is equivalent to

$$TZ^tx = 0 \quad \text{and} \quad Tx = ZTZ^tx.$$

Thus $x \in K$ if and only if

$$TZ^tx = 0 \quad \text{and} \quad Tx = 0. \tag{3.1}$$

Define $y = Jx$ and notice that $x = Jy$. Thus

$$Ty = JTJy = JTx = 0, \tag{3.2}$$

$$TZy = JTJJZ^tJy = JTZ^tx = 0. \tag{3.3}$$

Now it follows from the definition of the generalized inverse T^- that matrices T and T^- have the same null space. Hence from (3.2),

$$T^-y = 0,$$

and from (3.3),

$$T^-Zy = 0.$$

These two results prove statement (i) of Lemma 3.2, namely,

$$W \subset \operatorname{Ker} \Delta T^-.$$

Next, K and $W = J(K)$ have the same dimension since J is one to one. This proves (ii). Finally, result (3.2) immediately gives us the assertion (iii). □

Proof of Theorem 3.1. On one hand, the quadratic form associated with the matrix ΔT^- is null on the subspace V. On the other hand, the subspace W is included in the null space of the linear operator ΔT^-. From these properties, it is straightforward to conclude that the quadratic form is null on the whole subspace $E = V + W$. Moreover, $V \perp W$ because $V \subset \operatorname{range} T$ and $W \subset \operatorname{Ker} T$, hence $E = V \oplus W$ and

$$\dim (E) = \dim (V) + \dim (W),$$

$$\dim (E) = \dim (V) + \dim (K).$$

The dimension of $V = TZ^t(\operatorname{Ker} \nabla T)$ is obtained from the dimension rule for the range and null space of the restriction of TZ^t to $\operatorname{Ker} \nabla T$:

$$\dim (V) = \dim (\operatorname{Ker} \nabla T) - \dim (K). \tag{3.4}$$

Thus

$$\dim (E) = \dim (\operatorname{Ker} \nabla T) = N - r.$$

The quadratic form $\langle \Delta T^{-}x, x\rangle$ vanishes at least on a subspace E of dimension $N - r$. Consider an orthogonal basis whose first $N - r$ vectors form a basis of E. In such a basis this quadratic form is defined by a matrix $P\Delta T^{-}P^{t}$ that has at most r nonzero rows and r nonzero columns, as shown in the following:

$$P\Delta T^{-}P^{t} = \begin{pmatrix} & & & & & X \\ & & & & & X \\ & & 0 & & & X \\ & & & & & X \\ & & & & & X \\ X & X & X & X & X & X \end{pmatrix} \}\, r \text{ rows} .$$

Thus its rank is at most $2r$, and consequently the rank of ΔT^{-} is at most $2r$. □

Notice that Toeplitz matrices are persymmetric, and we therefore have Theorem 3.2.

THEOREM 3.2. *Let T be an $N \times N$ symmetric Toeplitz matrix, and Z be given as in Definition* 3.2. *Then the displacement rank of its generalized inverse T^{-} is bounded by* 4.

Remark 3.1. In order to generate a positive Toeplitz matrix of rank r for numerical experiments, we use a form of Caratheodory's representation [13], constructed with the help of Krylov subspaces. We first build an arbitrary orthogonal matrix Q with an invariant subspace of dimension N-r, as a product of r arbitrarily chosen symmetries. Then, we generate a vector x at random and we form the matrix M, whose columns are the vectors x, $Qx, \ldots, Q^{N-1}x$. Notice that the Krylov subspace built with matrix Q and starting vector x is indeed of dimension r. The required matrix is the covariance matrix $T = M^{t}M$.

Remark 3.2. As particular cases, let us insist that generalized inverses of $N \times N$ symmetric Toeplitz matrices of rank 1 have in general a displacement rank of 2. If their rank is larger than 1 and smaller than N, they will have in general a displacement rank of 4 from Theorem 3.2. In particular cases, however, their displacement rank may fall to 3 (see the example below).

Example. This simple example illustrates the practical issues addressed in the remarks above. Let $N = 5$ and $r = 3$, and define the vectors $a = (1\ 0\ 0\ 0\ 0)^{t}$, $b = (1/\sqrt{2}\ 1/\sqrt{2}\ 0\ 0\ 0)^{t}$, and $c = (0\ 1/\sqrt{2}\ 1/\sqrt{2}\ 0\ 0)^{t}$. Following the procedure proposed in Remark 3.1, the orthogonal matrix $Q = (I - 2bb^{t})(I - 2cc^{t})$ is built, and the matrix $M = [a\ Qa\ Q^{2}a\ Q^{3}a\ Q^{4}a]$,

$$Q = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad M = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Now, a symmetric Toeplitz matrix of rank 3 can be obtained by computing $M^{t}M$. We get

$$T = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \qquad 4T^{-} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 4 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

It can be checked that the displacement rank of T^- is 3. Another example would show that the bound can be reached for the same value of rank (T). Change a into $(0\ 1\ 0\ 1\ 0)^t$, for instance, and get a displacement rank of 4 for T^-, with

$$T=\begin{pmatrix}2&1&1&2&1\\1&2&1&1&2\\1&1&2&1&1\\2&1&1&2&1\\1&2&1&1&2\end{pmatrix},\qquad 16T^-=\begin{pmatrix}3&-1&-2&3&-1\\-1&3&-2&-1&3\\-2&-2&12&-2&-2\\3&-1&-2&3&-1\\-1&3&-2&-1&3\end{pmatrix}.$$

4. Nonsymmetric case. In this section we extend Theorem 3.1 to the wider class of persymmetric matrices, containing in particular the nonsymmetric Toeplitz matrices.

THEOREM 4.1. *Let T be an $N\times N$ persymmetric matrix with displacement rank r. Then the displacement rank of its generalized inverse T^- is bounded by $2r$.*

First we need to modify Lemma 3.1.

LEMMA 4.1. *The subspaces $V_1 = TZ^t(\text{Ker } \nabla T)$ and $V_2 = T^tZ^t(\text{Ker } \nabla T^t)$ are orthogonal for the bilinear form $\langle \Delta T^- y, z\rangle$.*

Proof. The proof is derived in a similar manner as in Lemma 3.1. Let $v_1 = TZ^tu_1$ with u_1 in Ker ∇T and let $v_2 = T^tZ^tu_2$ with u_2 in Ker ∇T^t. Now calculate $\langle \Delta T^-v_1, v_2\rangle$,

$$\langle \Delta T^-v_1, v_2\rangle = \langle T^-TZ^tu_1, T^tZ^tu_2\rangle - \langle Z^tT^-ZTZ^tu_1, T^tZ^tu_2\rangle.$$

By using the transpose rule for operators, we get

$$\langle \Delta T^-v_1, v_2\rangle = \langle TT^-TZ^tu_1, Z^tu_2\rangle - \langle T^-Tu_1, ZT^tZ^tu_2\rangle.$$

But $ZT^tZ^tu_2 = T^tu_2$ for $u_2 \in$ Ker ∇T^t, and $TT^-T = T$ by definition of T^-. Hence

$$\langle \Delta T^-v_1, v_2\rangle = \langle TZ^tu_1, Z^tu_2\rangle - \langle T^-Tu_1, T^tu_2\rangle,$$

and resorting to transposition gives

$$\langle \Delta T^-v_1, v_2\rangle = \langle ZTZ^tu_1, u_2\rangle - \langle TT^-Tu_1, u_2\rangle.$$

Another use of the properties $ZTZ^tu_1 = Tu_1$ for $u_1 \in$ Ker ∇T and $TT^-T = T$ yields

$$\langle \Delta T^-v_1, v_2\rangle = \langle Tu_1, u_2\rangle - \langle Tu_1, u_2\rangle = 0. \qquad \square$$

Next we modify Lemma 3.2.

DEFINITION 4.1. Define K_1 as the null space of the operator TZ^t in Ker ∇T, i.e.,

$$K_1 = \text{Ker } TZ^t \cap \text{Ker } \nabla T,$$

and define K_2 as the null space of the operator T^tZ^t in Ker ∇T^t, i.e.,

$$K_2 = \text{Ker } T^tZ^t \cap \text{Ker } \nabla T^t.$$

Also define W_i as the subspace $J(K_i)$.

LEMMA 4.2. *The subspace W_1 is included in the null space of the linear operator ΔT^-, and W_2 is included in the null space of the linear operator ΔT^{t-}. Moreover,*

$$\dim(W_i) = \dim(K_i),$$

$$W_1 \subset \text{Ker } T^t \quad \textit{and} \quad W_2 \subset \text{Ker } T.$$

Proof. Recall that $T^t = JTJ$, $Z = JZ^tJ$, and $J^2 = I_N$. First, let us check that W_1 is included in Ker ΔT^-. Let x be in K_1 and remark that this is equivalent to

$$TZ^tx = 0 \quad \text{and} \quad Tx = ZTZ^tx.$$

Thus $x \in K_1$ if and only if

$$TZ^t x = 0 \quad \text{and} \quad Tx = 0. \tag{4.1}$$

Define $y = Jx$. Noticing that $x = Jy$ gives

$$T^t y = JTJy = JTx = 0, \tag{4.2}$$

$$T^t Zy = JTJJZ^t Jy = JTZ^t x = 0. \tag{4.3}$$

Now by definition of the generalized inverse T^-, the operators T^t and T^- have the same null space. Hence

$$T^- y = 0 \quad \text{and} \quad T^- Zy = 0.$$

This proves that $W_1 \subset \operatorname{Ker} \Delta T^-$. Next, J is one to one; thus K_1 and $W_1 = J(K_1)$ have the same dimension, whereas the result (4.2) gives us $W_1 \subset \operatorname{Ker} T^t$. A similar proof holds for W_2. □

Proof of Theorem 4.1. The subspaces V_1 and V_2 are orthogonal with respect to the bilinear form associated with the matrix ΔT^-. The subspace W_1 is included in the null space of the linear operator ΔT^- while the subspace W_2 is included in the null space of the linear operator ΔT^{t-}. From these properties, it is straightforward to conclude that the whole subspaces $E_1 = V_1 + W_1$ and $E_2 = V_2 + W_2$ are orthogonal, according to the bilinear form associated with the matrix ΔT^-. Indeed, if y_1, z_1, y_2, and z_2 are, respectively, in V_1, W_1, V_2, and W_2,

$$\langle \Delta T^-(y_1 + z_1), (y_2 + z_2) \rangle = \langle \Delta T^- y_1, y_2 \rangle + \langle \Delta T^- z_1, (y_2 + z_2) \rangle + \langle y_1, \Delta T^{t-} z_2 \rangle,$$

and each term on the right-hand side of this expression is null from preceding results.

Moreover, $V_1 \perp W_1$ because $V_1 \subset$ range T and $W_1 \subset \operatorname{Ker} T^t$. Hence $E_1 = V_1 \oplus W_1$ and, following the proof of the symmetric case, we may conclude that

$$\dim (E_1) = N - r.$$

The bilinear form $\langle \Delta T^- x, y \rangle$ has two orthogonal subspaces E_1 and E_2, both of dimension at least $N - r$. It is easy to conclude, as in the symmetric case, that the rank of this bilinear form is at most $2r$ and consequently the rank of matrix ΔT^- is at most $2r$. □

5. Concluding remarks. Theorem 4.1 says that if a persymmetric matrix T has a displacement rank r with respect to a persymmetric displacement matrix Z, then its pseudo-inverse T^- has a displacement rank bounded by $2r$. In other words, the matrices T and T^- can be completely characterized by at most $2r$ and $4r$ generating vectors, respectively. If T is symmetric Toeplitz, four vectors are sufficient to characterize matrix T^-. Now in order for this result to be fully exploited, it would be necessary to find an algorithm able to express explicitly the generators of T^-. This issue is left open.

Acknowledgment. The authors are indebted to the pertinent recommendations of the referees.

REFERENCES

[1] T. KAILATH, *Signal processing applications of some moment problems*, in Proceedings of Symposia in Applied Mathematics, Vol. 37, San Antonio, TX, January 1987, American Mathematical Society, Providence, RI, 1987, pp. 71–109.

[2] T. KAILATH, S. Y. KUNG, AND M. MORF, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[3] B. FRIEDLANDER, M. MORF, T. KAILATH, AND L. LJUNG, *New inversion formulas for matrices classified in terms of their distance from Toeplitz matrices*, Linear Algebra Appl., 27 (1979), pp. 31–60.

[4] C. GUEGUEN, *An introduction to displacement ranks*, in Signal Processing, Vol. XLV, J. L. Lacoume, T. S. Durrani, R. Stora, eds., Elsevier, New York, 1987, pp. 705–780.

[5] J. CHUN, *Fast array algorithms for structured matrices*, Ph.D. thesis, Stanford Univ., Stanford, CA, June 1989.

[6] Y. GENIN, P. VAN DOOREN, T. KAILATH, J. M. DELOSME, AND M. MORF, *On Σ-lossless transfer functions*, Linear Algebra Appl., 50 (1983), pp. 251–275.

[7] H. LEV-ARI AND T. KAILATH, *Lattice filter parametrization and modeling of nonstationary processes*, IEEE Trans. Inform. Theory, 30 (1984), pp. 2–16.

[8] G. H. GOLUB, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[9] G. S. AMMAR AND W. B. GRAGG, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[10] T. KAILATH, A. VIEIRA, AND M. MORF, *Inverses of Toeplitz operators, innovations and orthogonal polynomials*, SIAM Rev., 20 (1978), pp. 106–119.

[11] I. S. IOHVIDOV, *Hankel and Toeplitz Matrices and Forms: Algebraic Theory*, Birkhauser, Boston, 1982.

[12] P. DELSARTE, Y. GENIN, AND Y. G. KAMP, *A generalization of the Levinson algorithm for Hermitian Toeplitz matrices with any rank profile*, IEEE Trans. Acoust. Speech Signal Process, 33 (1985), pp. 964–971.

[13] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, CA, 1958.

[14] S. POMBRA, H. LEV-ARI, AND T. KAILATH, *Levinson and Schur algorithms for Toeplitz matrices with singular minors*, Internat. Conf. ICASSP, New York, April 1988, pp. 1643–1646.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 655–669, July 1993

004

SPARSITY ANALYSIS OF THE QR FACTORIZATION*

DONOVAN R. HARE†, CHARLES R. JOHNSON‡, D. D. OLESKY¶,
AND P. VAN DEN DRIESSCHE†§

Abstract. Given only the zero-nonzero pattern of an $m \times n$ matrix A of full column rank, which entries of Q and which entries of R in its QR factorization must be zero, and which entries may be nonzero? A complete answer to this question is given, which involves an interesting interplay between combinatorial structure and the algebra implicit in orthogonality. To this end some new sparse structural concepts are introduced, and an algorithm to determine the structure of Q is given. The structure of R then follows immediately from that of Q and A. The computable zero/nonzero structures for the matrices Q and R are proven to be tight, and the conditions on the pattern for A are the weakest possible (namely, that it allows matrices A with full column rank). This complements existing work that focussed upon R and then only under an additional combinatorial assumption (the strong Hall property).

Key words. QR factorization, bipartite graph, Hall property, sparse matrix, combinatorial matrix theory, orthogonality

AMS subject classifications. 05C50, 15A23, 65F25, 65F50

1. Introduction.

For $m \geqq n$, an $m \times n$ matrix $A = [a_{ij}]$ has a unique factorization $A = QR$, in which $Q = [q_{ij}]$ is $m \times n$ and has orthonormal columns and $R = [r_{ij}]$ is $n \times n$ and upper triangular with positive diagonal entries, exactly when A has full rank n. In this event, we say that A has a unique QR factorization, and we mean the one in which R has positive diagonal. There are several possible ways to compute such a factorization (with exact arithmetic), for example, using plane rotations (method of Givens), Householder transformations, or the Gram–Schmidt procedure. When A has a unique QR factorization, it does not matter, except for insight or computational considerations, which algorithm is used to obtain the factorization.

By a *nonzero pattern* $\mathscr{A}$ of size $m \times n$ we mean the set of positions in an $m \times n$ matrix in which the nonzero entries occur. We typically describe such a pattern via an $m \times n$ array with 0 and $*$ entries, in which $*$ denotes a nonzero. For example,

$$A = \begin{bmatrix} -3 & 0 \\ 1 & 2 \\ 0 & \pi \end{bmatrix}$$

is a matrix with pattern

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

* Received by the editors March 26, 1990; accepted for publication (in revised form) November 21, 1991.

† Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada V8W 3P4 (dhare@oksun1.okanagan.bc.ca) and (pvdd@uvvm.uvic.ca).

‡ The work of this author was supported in part by National Science Foundation grant DMS90-00839 and Office of Naval Research contract N00014-90-J-1739. Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23185 (crjohn@wmvm1.bitnet).

¶ The work of this author was supported in part by Natural Sciences and Engineering Research Council of Canada grant A-8214 and by the University of Victoria Committee on Faculty Research and Travel. Department of Computer Science, University of Victoria, Victoria, British Columbia, Canada V8W 3P6 (dolesky@uvunix.uvic.ca).

§ The work of this author was supported in part by Natural Sciences and Engineering Research Council of Canada grant A-8965 and by the University of Victoria Committee on Faculty Research and Travel.

If matrix A has pattern $\mathscr{A}$, then we write $A \in \mathscr{A}$. By $\mathscr{A} + \mathscr{B}$ (or $\mathscr{A} \cup \mathscr{B}$) we mean the *union* of two patterns of the same size, that is, a pattern that has a zero in a given position only when both patterns are zero in that position. Multiplication of patterns is carried out under the following assumptions: $(* \cdot *) = *$ and $(* \cdot 0) = (0 \cdot *) = (0 \cdot 0) = 0$. Two column vectors (or vector patterns) are *combinatorially orthogonal* if each term in the inner product is zero.

Our primary purpose is to give a complete solution to the following problem. For $m \geqq n$, given an $m \times n$ nonzero pattern $\mathscr{A}$ that allows full rank, determine the union, over all full column rank matrices $A \in \mathscr{A}$, of all patterns occurring in the matrix Q and in the matrix R such that $A = QR$ is the QR factorization of A. The resulting unions are denoted by $\mathscr{Q}$ and $\mathscr{R}$, respectively. Our solution to this problem continues earlier work ([GH], [GLN], [CEG], and a small part of [GN]), which focussed mainly on the upper triangular pattern $\mathscr{R}$ (rather than $\mathscr{Q}$) and gave a description only under an additional combinatorial assumption (the strong Hall property) on the pattern $\mathscr{A}$. The present work is based on the Gram–Schmidt procedure and is in the spirit of an analogous combinatorial analysis of the LU factorization of a square matrix that was given in [JOD]. However, there are subtleties that are quite different from the LU case.

We note that, for a given pattern $\mathscr{A}$, different zero patterns may occur among the Qs, or among the Rs, in the QR factorization of different full rank matrices A with pattern $\mathscr{A}$. This is due to the possibility of chance numeric relations among the entries of a specific matrix A, which cause "accidental zeros." Our interest is in ignoring chance cancellation and in obtaining the union of all patterns for both Q and R. Knowledge of $\mathscr{Q}$ and $\mathscr{R}$ is useful for computational purposes, as they are the smallest patterns that are guaranteed to contain the nonzero entries of the QR factorization of an arbitrary $A \in \mathscr{A}$. Thus, $\mathscr{Q}$ and $\mathscr{R}$ could be used to define a suitable data structure for Q and R, respectively.

Some particularly simple cases occur when $\mathscr{A}$ is square. If $\mathscr{A}$ is full upper triangular, then $\mathscr{R}$ has the same pattern and $\mathscr{Q}$ is diagonal. If $\mathscr{A}$ is full upper Hessenberg, then $\mathscr{Q}$ has this same pattern and $\mathscr{R}$ is full. However, if $\mathscr{A}$ is full lower triangular, then both $\mathscr{R}$ and $\mathscr{Q}$ are full.

A combinatorial necessary condition for an $m \times n$ matrix A to have rank n is the following (see, e.g., [CEG]).

DEFINITION. For $m \geqq n$, an $m \times n$ matrix A (or pattern $\mathscr{A}$) satisfies the *Hall property* if every k columns, $1 \leqq k \leqq n$, collectively have nonzero entries in at least k rows.

It is clear that there are full rank matrices with pattern $\mathscr{A}$ if and only if $\mathscr{A}$ satisfies the Hall property. So that our problem is well defined, we wish to assume that each matrix A to be factored has a unique QR factorization. Thus the problem statement assumed that A has full column rank, and therefore, we restrict our attention to patterns satisfying the Hall property. If A has a unique QR factorization $A = QR$, then, for an $m \times m$ permutation matrix P, PA also has a unique QR factorization $PA = (PQ)R$. Since Q is modified in this simple way, and R not at all, we may permute the rows of A for convenience. For example, under the Hall property assumption, we may permute the rows so that the diagonal entries are all nonzero. We note, however, that permutation of the columns of A, in general, radically changes both Q and R.

In [CEG] the possible nonzeros for $\mathscr{R}$ are identified under an additional assumption. (The following definition, a slight variant from what appears in [CEG], now seems to be standard.)

DEFINITION. For $m \geqq n$, an $m \times n$ matrix A (or pattern $\mathscr{A}$) satisfies the *strong Hall property* if

(i) $m = n > 1$ and every k columns, $1 \leqq k < n$, collectively have nonzero entries in *more* than k rows, or

(ii) $m > n$ and every k columns, $1 \leqq k \leqq n$, collectively have nonzero entries in *more* than k rows.

Note that, if $m = n$ and each main diagonal entry of A is nonzero, then A satisfies the strong Hall property if and only if A is irreducible. Under the strong Hall assumption on $\mathscr{A}$ it is shown in [CEG] that $\mathscr{R}$ is accurately predicted by applying "symbolic Gaussian elimination" (i.e., "symbolic LU factorization") to the symbolic product $\mathscr{A}^T\mathscr{A}$, or by doing "symbolic Givens rotations" on $\mathscr{A}$. In this case, $\mathscr{R}$ can also be bounded by the pattern of U in the symbolic LU factorization (with partial pivoting) of $\mathscr{A}$; see [GN].

Without the strong Hall assumption, the complete analysis is considerably more delicate. There can be interplay between the combinatorics of the arrangement of zeros and nonzeros and the algebra (associated with orthogonality required in Q) of the QR calculation that forces zeros that are missed by any naive analysis. This is illustrated in the following example from [CEG], in which $\mathscr{A}$ satisfies the Hall property but not the strong Hall property.

Example 1.1. Let

$$\mathscr{A} = \begin{bmatrix} * & 0 & 0 & 0 & 0 & 0 \\ * & * & * & 0 & 0 & 0 \\ 0 & 0 & * & * & * & * \\ 0 & 0 & 0 & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & 0 \\ * & 0 & 0 & 0 & 0 & * \end{bmatrix}. \tag{1.1}$$

As pointed out in [Ge], when $\mathscr{Q}$ and $\mathscr{R}$ are predicted by a symbolic analog of the (numeric) method of Givens (where nonzero entries in the jth column are zeroed out in the order $a_{j+1,j}$ to $a_{n,j}$), too much fill is predicted. Specifically, this approach gives

$$\mathscr{Q} = \begin{bmatrix} * & * & \otimes & \otimes & \otimes & * \\ * & * & \otimes & \otimes & \otimes & \otimes \\ 0 & 0 & * & \otimes & \otimes & \otimes \\ 0 & 0 & 0 & * & \otimes & \otimes \\ 0 & 0 & 0 & 0 & * & \otimes \\ * & * & \otimes & \otimes & \otimes & * \end{bmatrix}, \qquad \mathscr{R} = \begin{bmatrix} * & * & * & 0 & 0 & * \\ & * & * & 0 & 0 & * \\ & & * & * & * & * \\ & & & * & \otimes & \otimes \\ 0 & & & & * & \otimes \\ & & & & & * \end{bmatrix}, \tag{1.2}$$

where $\otimes = *$, whereas $\otimes = 0$ gives the true patterns. The entries in $\mathscr{R}$ designated by $\otimes$ are called *bogus fill* in [CEG]. The pattern for $\mathscr{R}$ in (1.2) with $\otimes = *$ is also obtained when symbolic LU factorization of the symbolic product $\mathscr{A}^T\mathscr{A}$ is performed [GH] or when symbolic LU factorization (with partial pivoting) of $\mathscr{A}$ is performed [GN].

In the next section we begin by defining the notion of a Hall set. This enables us to define a sequence of bipartite graphs that lead to the identification of all zero entries in $\mathscr{Q}$. The orthogonality of Q can require zero entries in $\mathscr{Q}$ in somewhat subtle ways. Some graph-theoretic results concerning these bipartite graphs are given in § 3. The purpose of § 4 is to show that the remaining entries of $\mathscr{Q}$ are, in fact, nonzeros. It begins with explicit consideration of a special form for the pattern of $\mathscr{A}$ and then uses embedding and continuity to prove our result for general patterns. An algorithm for determining $\mathscr{Q}$ is given. In § 5 it is shown how to determine the pattern $\mathscr{R}$ in a simple way, once the pattern $\mathscr{Q}$ is known.

2. Zero entries in $\mathscr{Q}$. Related to the Hall properties, we introduce the concept of a Hall set. Given any $m \times n$ matrix A (or pattern $\mathscr{A}$) we denote the jth column of A (or $\mathscr{A}$) by A_j (or $\mathscr{A}_j$), and the restriction of A_j to the row index set α by $A_j[\alpha]$.

DEFINITION. For an $m \times n$ matrix A (or pattern $\mathscr{A}$) with $m \geqq n$, a set of k columns in A (or $\mathscr{A}$), $1 \leqq k \leqq n$ is a *Hall set* if these columns collectively have nonzero entries in *exactly* k rows.

Under the assumption that A (or $\mathscr{A}$) has the Hall property, the union of two Hall sets is a Hall set, so there exists a unique Hall set of maximum cardinality ($\geqq 0$) in any given set of columns. Let S_j be the *Hall set of maximum cardinality* in the first j columns; we define $S_0 = \phi$. For example, in the pattern (1.1), $S_1 = \phi$ and $S_j = \{\mathscr{A}_2, \ldots, \mathscr{A}_j\}$ for $2 \leqq j \leqq 5$. Note that if an $m \times n$ matrix has the strong Hall property, then $S_j = \phi$, $1 \leqq j \leqq n - 1$.

The Gram–Schmidt orthonormalization process (see, e.g., [HJ, § 0.6.4]) can be used to construct the matrix Q of the QR factorization of a matrix A. We show below how to adapt the Gram–Schmidt procedure to determine $\mathscr{Q}$ for a pattern $\mathscr{A}$. Clearly $\mathscr{Q}_1 = \mathscr{A}_1$ as Q_1 is a (nonzero) multiple of A_1. If $\mathscr{A}_2$ is combinatorially orthogonal to $\mathscr{A}_1$, then $\mathscr{Q}_2 = \mathscr{A}_2$. However, if $\mathscr{A}_2$ is not combinatorially orthogonal to $\mathscr{A}_1$ and $S_1 = \phi$, then $\mathscr{Q}_2 = \mathscr{A}_1 + \mathscr{A}_2 = \mathscr{Q}_1 + \mathscr{A}_2$. The effect of Hall sets on this columnwise "naive accumulation" of nonzero entries is now investigated.

Let s_j be the set of all row indices covered by the columns of S_j; thus $|S_j| = |s_j|$. (Note that if we assume all diagonal entries of $\mathscr{A}$ are $*$, then $s_j = \{i\colon 1 \leqq i \leqq j, \mathscr{A}_i \in S_j\}$; however, we do not make this assumption in this section.) For a given $\mathscr{A}$, we define a bipartite graph that leads to a partition of the row indices of $\mathscr{A}$ and identifies the zero entries in a column of $\mathscr{Q}$. For a fixed j, $1 \leqq j \leqq n$, the bipartite graph $B_j(\mathscr{A}) = (R_j(\mathscr{A}), C_j(\mathscr{A}); E_j(\mathscr{A}))$ is defined as follows:

$$C_j(\mathscr{A}) = \{c_k = k\colon 1 \leqq k \leqq j \text{ and } \mathscr{A}_k \notin S_{j-1}\};$$

$$R_j(\mathscr{A}) = \{r_i = i\colon 1 \leqq i \leqq m \text{ and there exists } c_k \in C_j(\mathscr{A}) \text{ with a } * \text{ in row } r_i \text{ and } r_i \notin s_{j-1}\};$$

$$\{r_i, c_k\} \in E_j(\mathscr{A}), \text{ the undirected edge set, if and only if } r_i \in R_j(\mathscr{A}) \text{ and } c_k \in C_j(\mathscr{A}) \text{ and the } (i, k) \text{ position of } \mathscr{A} \text{ is } *.$$

Thus, to construct $B_j(\mathscr{A})$, delete columns in S_{j-1}, rows s_{j-1} and zero rows from $[\mathscr{A}_1, \ldots, \mathscr{A}_j]$, and take the bipartite graph of the remaining pattern (with the labelling inherited from the original pattern).

A *path* (undirected) in $B_j(\mathscr{A})$ is an alternating sequence of row and column vertices of the form $r_{k_1}, c_{j_1}, r_{k_2}, \ldots, r_{k_i}, c_{j_i}$ with each edge $\{r_{k_1}, c_{j_1}\}, \{c_{j_1}, r_{k_2}\}, \ldots, \{r_{k_i}, c_{j_i}\} \in E_j(\mathscr{A})$. Such a path is associated with the sequence $(k_1, j_1), (k_2, j_1), \ldots, (k_i, j_i)$ of entries that are $*$ in $\mathscr{A}$. A path may begin or end with either a row or column vertex, may have just one vertex, and is always assumed to be simple.

Let p_j be the set of all row indices occurring in $R_j(\mathscr{A})$ that can be reached by a path from c_j. Let u_j be the set of all row indices occurring in $R_j(\mathscr{A})$ that cannot be reached by a path from c_j. Finally, let t_j be the set of all row indices of $\mathscr{A}$ not in $R_j(\mathscr{A})$ or s_{j-1}. We then have that s_{j-1}, p_j, u_j, and t_j form a partition of all m row indices.

To illustrate this partition, consider again the pattern $\mathscr{A}$ in example (1.1); the bipartite graphs for $j = 3$ and for $j = 6$ are shown in Fig. 1.

We use the partition defined above to identify the zero entries in a column of $\mathscr{Q}$.

THEOREM 2.1. *For $m \geqq n$, let $\mathscr{A}$ be an $m \times n$ pattern with the Hall property. For any fixed j, $1 \leqq j \leqq n$, the entries of $\mathscr{Q}_j$ in positions s_{j-1}, t_j, and u_j are zero.*

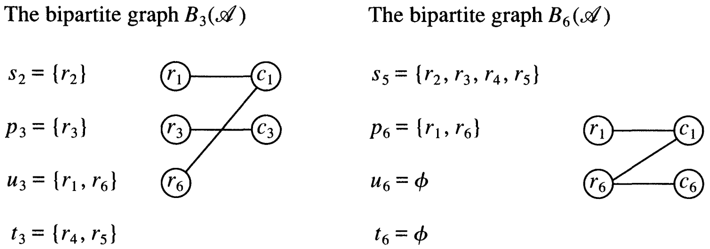


FIG. 1

Proof. We rely on two facts that follow from the Gram–Schmidt procedure:

(i) $Q_j \in \text{span}\,\{A_1, \dots, A_j\}$, and

(ii) Q_j is orthogonal to $A_1, \dots, A_{j-1}$.

Consider first the positions t_j. In these rows the entries in columns $A_1, \dots, A_j$ of any $A \in \mathscr{A}$ are all zero; so by (i), the positions t_j are all zero for any $Q_j \in \mathscr{Q}_j$.

For the positions s_{j-1}, fact (ii) above implies that Q_j is orthogonal to each column in S_{j-1}. But the columns of S_{j-1}, restricted to rows s_{j-1}, span a subspace of dimension $|s_{j-1}|$ and are zero outside these rows (by the definition of a Hall set). Thus $Q_j[s_{j-1}]$ is orthogonal to every vector in the above subspace, and so must be zero. Thus the positions s_{j-1} of any $Q_j \in \mathscr{Q}_j$ are zero; these are due to the presence of a Hall set.

Finally consider the positions u_j, and let $\{A_{i_1}, \dots, A_{i_k}\}$ be the set of columns on a path in $B_j(\mathscr{A})$ from indices in u_j. From (i) and because $u_j \cap (p_j \cup s_{j-1}) = \phi$, $Q_j[u_j] \in \text{span}\,\{A_{i_1}[u_j], \dots, A_{i_k}[u_j]\}$. Because of (ii), and because $A_{i_g}[p_j] = 0$, $g = 1, \dots, k$, and $Q_j[s_{j-1}] = 0$ (and $A_{i_g}[t_j] = 0$, $g = 1, \dots, k$, and $Q_j[t_j] = 0$), $Q_j[u_j]$ is orthogonal to each $A_{i_g}[u_j]$. However, as the intersection of a subspace and its orthogonal complement is zero, $Q_j[u_j] = 0$ and the positions u_j of any $Q_j \in \mathscr{Q}_j$ are zero. □

We have thus identified three sets of row indices in $\mathscr{Q}_j$ that are zero for any full rank matrix $A \in \mathscr{A}$ with $A = QR$. In particular, the presence of a Hall set in columns $\{\mathscr{A}_1, \dots, \mathscr{A}_{j-1}\}$ means that $q_{i,j} = 0$ for $i \in s_{j-1}$ (in fact, $q_{i,j+k} = 0$ for $i \in s_{j-1}$ and $k = 0, \dots, n-j$). If $i \in t_j$, that is, $a_{ik} = 0$ for $k = 1, \dots, j$, then $q_{ij} = 0$. If $i \in u_j$, then $q_{ij} = 0$ because of combinatorial orthogonality in the columns of $\mathscr{A}$, which occurs in a subtle way (see Corollary 4.9).

3. Strong Hall bipartite graphs.

In this section we analyze the strong Hall property from a graph-theoretic point of view. The objective is to obtain a result (Thm. 3.2) concerning the bipartite graphs $B_j(\mathscr{A})$ defined in § 2. We first prove a preliminary result concerning strong Hall bipartite graphs, which we now define.

DEFINITION. Let $G(X, Y; E)$ denote a bipartite graph with vertex sets X, Y and undirected edge set E and if $S \subseteq Y$, let $N_G(S)$ denote the set of all neighbors in X of vertices in S. Then $G(X, Y; E)$ is *strong Hall* with respect to Y if

(i) $|X| = |Y| > 1$ and $|S| < |N_G(S)|$, for all proper nonempty subsets S of Y, or

(ii) $|X| > |Y|$ and $|S| < |N_G(S)|$, for all nonempty subsets S of Y.

There is an obvious equivalence between patterns $\mathscr{A}$ that satisfy the strong Hall property and bipartite graphs $G(X, Y; E)$ that are strong Hall with respect to Y when the vertex sets X, Y are associated with the row and column sets of $\mathscr{A}$, respectively, and edges correspond to nonzero entries in $\mathscr{A}$. Note that the bipartite graphs $B_j(\mathscr{A})$ defined in § 2 are strong Hall with respect to $C_j(\mathscr{A})\backslash\{c_j\}$.

The following theorem is used to prove Theorem 3.2, but is of independent interest as it gives a characterization of *strong Hall critical graphs* $G = (X, Y; E)$ with

$|X| > |Y|$ (i.e., strong Hall graphs that are no longer strong Hall if any one edge is removed). Let G be strong Hall with respect to Y. If for all $y \in Y$, the degree $d_G(y) = 2$, then it is clear that G is strong Hall critical with respect to Y. We now show that the converse is also true.

THEOREM 3.1. *Let $G = (X, Y; E)$ be strong Hall with respect to Y, where $|X| > |Y|$. If $y \in Y$ has degree $d_G(y) \geqq 3$, then there is at most one edge $e = \{x, y\} \in E$ such that $G - e$ is not strong Hall with respect to Y.*

Proof. Suppose there exists $y \in Y$ with $d_G(y) \geqq 3$, and there is an edge $e = \{x, y\} \in E$ such that $H = G - e$ is not strong Hall with respect to Y. Let $S \subseteq Y$, $S \neq \phi$, be such that $|N_H(S)| \leqq |S|$. If $y \notin S$, then $N_H(S) = N_G(S)$, and hence $|N_H(S)| = |N_G(S)| > |S|$ since G is strong Hall with respect to Y. Therefore, $y \in S$. If $x \in N_G(S\setminus\{y\})$, then again $N_H(S) = N_G(S)$, a contradiction, and so $x \notin N_G(S\setminus\{y\})$. Thus $|N_G(S)| = |N_H(S)| + 1 \leqq |S| + 1$, and since G is strong Hall with respect to Y, $|N_G(S)| = |S| + 1$.

Now suppose there exists $y \in Y$ with $d_G(y) \geqq 3$ and there exist distinct edges $e_1 = \{x_1, y\}$ and $e_2 = \{x_2, y\}$ in E such that for $i = 1, 2$, $H_i = G - e_i$ is not strong Hall with respect to Y. Then for $i = 1, 2$ there exists $S_i \subseteq Y$, $S_i \neq \phi$, such that $|N_{H_i}(S_i)| \leqq |S_i|$. From above, $y \in S_1 \cap S_2$, and for $i = 1, 2$,

$$|N_G(S_i)| = |S_i| + 1 \tag{3.1}$$

and

$$x_i \notin N_G(S_i\setminus\{y\}). \tag{3.2}$$

First suppose $(S_1 \cap S_2)\setminus\{y\} = S \neq \phi$. For each $z \in S$ and $i = 1, 2$, $\{z, x_i\}$ is not an edge in G by (3.2). Thus

$$\begin{aligned} |N_G(S_1 \cap S_2)| &= |N_G(\{y\})\setminus N_G(S)| + |N_G(S)| \\ &\geqq 2 + |N_G(S)| \quad \text{since } x_1, x_2 \notin N_G(S) \\ &> 2 + |S| \quad \text{since } S \neq \phi \quad \text{and} \quad G \text{ is strong Hall} \\ &= 2 + (|S_1 \cap S_2| - 1) \\ &= |S_1 \cap S_2| + 1. \end{aligned}$$

If, on the other hand, $S_1 \cap S_2 = \{y\}$, then

$$|N_G(S_1 \cap S_2)| = |N_G(\{y\})| \geqq 3 > |S_1 \cap S_2| + 1.$$

Therefore, in any case,

$$|N_G(S_1) \cap N_G(S_2)| \geqq |N_G(S_1 \cap S_2)| \geqq |S_1 \cap S_2| + 2. \tag{3.3}$$

Thus

$$\begin{aligned} |S_1 \cup S_2| &= (|S_1| + 1) + (|S_2| + 1) - (|S_1 \cap S_2| + 2) \\ &\geqq |N_G(S_1)| + |N_G(S_2)| - |N_G(S_1) \cap N_G(S_2)| \quad \text{by (3.1) and (3.3)} \\ &= |N_G(S_1) \cup N_G(S_2)| \\ &= |N_G(S_1 \cup S_2)|. \end{aligned}$$

This contradicts the stipulation that G be strong Hall with respect to Y (note that $|X| > |Y|$). Hence, at most one such edge e_i can exist. □

The next theorem provides a matching property of strong Hall bipartite graphs that is used to prove the main theorem in § 4. (Recall that a *matching* M in a bipartite graph G is a subset of its edges no two of which are adjacent. We say that M *saturates* a subset S of the vertices of G if every vertex in S is incident with one of the edges in M.) We will use the following notation in the next theorem. If γ_1 and γ_2 are paths in a graph G, and γ_1 ends in a vertex adjacent in G to the first vertex in γ_2, then $\gamma_1\gamma_2$ denotes the juxtaposition of γ_1 and γ_2 as alternating sequences of vertices and edges so that the edge from the end of γ_1 to the beginning of γ_2 is added giving a single path.

THEOREM 3.2. *Let $G = (X, Y; E)$ be a bipartite graph with $|X| \geqq |Y|$, let $y \in Y$, and suppose that $G - y$ is strong Hall with respect to $Y \setminus \{y\}$. If there is a path γ from y to x for some $x \in X$, then there exists a path τ from y to x and a matching M in $G - \tau$ such that M saturates all vertices in Y that are not on the path τ.*

Proof. Suppose the statement of the theorem is false, and let $G = (X, Y; E)$ be a counterexample with the fewest edges. Then there exists $x \in X$ and a path γ from y to x. Let $\hat{Y} = Y \setminus \{y\}$ and $\hat{G} = G - y$. Since G is a counterexample there must exist a vertex $\hat{y} \in Y$ not on γ.

If $d_{\hat{G}}(\hat{y}) \geqq 3$, then by Theorem 3.1, there exists $\hat{e} = \{\hat{x}, \hat{y}\}$ such that if $H = G - \hat{e}$, then $H - y$ is strong Hall with respect to $\hat{Y}$. Now γ is a path in H also, and since G is a counterexample with the fewest edges, there exists some path τ from y to x and a matching M in $H - \tau$ such that M saturates all the vertices in Y that are not on the path τ. But this matching and path are also in G, a contradiction.

Therefore, $d_{\hat{G}}(\hat{y}) = 2$ (note that $d_{\hat{G}}(\hat{y}) > 1$ since $\hat{G}$ is strong Hall with respect to $\hat{Y}$). Let x_1 and x_2 be the neighbors of $\hat{y}$ and let $H = (U, V; F)$ be the graph obtained from G by identifying x_1 and x_2 and by removing $\hat{y}$. That is, $U = (X \setminus \{x_1, x_2\}) \cup \{x^*\}$, where $x^* \notin X \cup Y$ and $V = Y \setminus \{\hat{y}\}$. Moreover, for all $u \in U$ and $v \in V$, $\{u, v\} \in F$ if and only if

(i) $u \in X \setminus \{x_1, x_2\}$, $v \in V$ and $\{u, v\} \in E$; or

(ii) $u = x^*$, $v \in V$, and $\{x_1, v\} \in E$ or $\{x_2, v\} \in E$.

Let $\hat{V} = V \setminus \{y\}$, $\hat{H} = H - y$, and note that since $|X| > |\hat{Y}|$, we have $|U| > |\hat{V}|$.

Claim. $\hat{H}$ is strong Hall with respect to $\hat{V}$.

Proof of claim. Let $S \subseteq \hat{V}$, $S \neq \phi$. If $x^* \notin N_{\hat{H}}(S)$, then $x_1 \notin N_{\hat{G}}(S)$ and $x_2 \notin N_{\hat{G}}(S)$. Thus $N_{\hat{H}}(S) = N_{\hat{G}}(S)$, and since $\hat{G}$ is strong Hall with respect to $\hat{Y}$, $|S| < |N_{\hat{G}}(S)| = |N_{\hat{H}}(S)|$.

Suppose there exists $S \subseteq \hat{V}$ such that $|S| \geqq |N_{\hat{H}}(S)|$. Then $x^* \in N_{\hat{H}}(S)$, by the above, and hence at least one of x_1 or x_2 is contained in $N_{\hat{G}}(S)$. If only one is contained in $N_{\hat{G}}(S)$, then since it is only replaced with x^* in H, we have $|N_{\hat{G}}(S)| = |N_{\hat{H}}(S)|$, a contradiction to $|S| < |N_{\hat{G}}(S)|$. Thus both x_1 and x_2 are contained in $N_{\hat{G}}(S)$. Since x_1 and x_2 are replaced with x^* in H, $|N_{\hat{G}}(S)| = |N_{\hat{H}}(S)| + 1$ and $N_{\hat{G}}(S \cup \{\hat{y}\}) = N_{\hat{G}}(S)$. But then

$$|S \cup \{\hat{y}\}| = |S| + 1 \geqq |N_{\hat{H}}(S)| + 1 = |N_{\hat{G}}(S)| = |N_{\hat{G}}(S \cup \{\hat{y}\})|$$

contradicts the fact that $\hat{G}$ is strong Hall with respect to $\hat{Y}$ (note that $S \cup \{\hat{y}\} \subseteq \hat{Y}$). This completes the proof of the claim.

We now identify a path η in H that is dependent upon the path γ in G. Table 1 describes how η is defined in H given the structure of γ. In Table 1, π, ρ, and σ are nonempty paths (i.e., they each contain at least one vertex).

Let η end at vertex u. Note that $u \in U$ and that η starts at vertex y. Since $\hat{H}$ is strong Hall with respect to $\hat{V}$, and H has fewer edges than G, there exists a path ω in H from y to u and matching $\hat{M}$ in $H - \omega$ that saturates all vertices in V that are not on ω. We will use this path and matching to construct similar ones in G. The constructions will differ

TABLE 1

Case	γ	Path η in H
1	γ (x_1 and x_2 not on γ)	γ
2	$\pi x_1 \rho$ or $\pi x_2 \rho$	$\pi x^* \rho$
3	$\pi x_1 \rho x_2 \sigma$ or $\pi x_2 \rho x_1 \sigma$	$\pi x^* \sigma$
4	$\pi x_1 \rho x_2$ or $\pi x_2 \rho x_1$	πx^*
5	πx_1 or πx_2	πx^*

depending on the structure of γ in G, but each will contradict the assumption that G is a counterexample to the theorem. The cases below are described in Table 1.

Cases 1–3. Here we have $x_1 \neq u \neq x_2$, and hence $u = x$.

If x^* is not on ω and x^* is not saturated by $\hat{M}$, then ω is a path in G from y to x and $M = \hat{M} \cup \{\{x_1, \hat{y}\}\}$ is a matching in G, which saturates all vertices in Y that are not on ω, a contradiction.

If there exists a $v \in V$ such that $\{x^*, v\} \in \hat{M}$, then x^* is not on ω, and hence, ω is a path in G from y to x. Moreover, $\{z, v\} \in E$ for some $z \in \{x_1, x_2\}$. Let $\{w\} = \{x_1, x_2\} \setminus \{z\}$. Thus $M = (\hat{M} \setminus \{\{x^*, v\}\}) \cup \{\{z, v\}, \{w, \hat{y}\}\}$ is a matching in G, which saturates all vertices in Y that are not on ω, a contradiction.

Therefore, x^* is on ω, and $\hat{M}$ is a matching in G. Let $\omega = \omega_1 x^* \omega_2$ and let v_1 and v_2 be the neighbors of x^* on ω such that ω_1 ends at v_1 and ω_2 starts at v_2. If $\{x_1, x_2\} \subseteq N_G(\{v_1, v_2\})$, then at least one of $\tau_1 = \omega_1 x_1 \hat{y} x_2 \omega_2$ and $\tau_2 = \omega_1 x_2 \hat{y} x_1 \omega_2$ is a path in G that starts at y and ends at x. Thus, for some $i \in \{1, 2\}$, $\hat{M}$ saturates all vertices in Y that are not on τ_i, a contradiction. Hence, $\{x_1, x_2\} \nsubseteq N_G(\{v_1, v_2\})$. Thus, since x^* is adjacent to v_1 and v_2, they are both adjacent to x_1 or to x_2 in G. If v_1 and v_2 are neighbors of x_1, then $\tau = \omega_1 x_1 \omega_2$ is a path in G from y to x, and $M = \hat{M} \cup \{\{x_2, \hat{y}\}\}$ is a matching in G that saturates all vertices in Y that are not on τ, a contradiction. Similarly, if v_1 and v_2 are neighbors of x_2, then $\tau = \omega_1 x_2 \omega_2$ is a path in G from y to x, and $M = \hat{M} \cup \{\{x_1, \hat{y}\}\}$ is a matching in G that saturates all vertices in Y that are not on τ, another contradiction. Therefore, Cases 1–3 cannot occur.

Cases 4 *and* 5. Note that $u = x^*$, and hence ω ends at x^*. Moreover, $\hat{M}$ is a matching in G. Let $\omega = \omega_1 x^*$ and let v_1 be the last vertex of the path ω_1. Since v_1 is adjacent to x^* in H, it is adjacent to either x_1 or x_2 in G. Thus, either $\tau_1 = \omega_1 x_1$, $\tau_2 = \omega_1 x_2 \hat{y} x_1$, $\tau_3 = \omega_1 x_2$, or $\tau_4 = \omega_1 x_1 \hat{y} x_2$ is a path from y to x in G. If τ_2 is such a path in G, then $\hat{M}$ saturates all vertices in Y that are not on τ_2; similarly, if τ_4 is a path from y to x in G. Otherwise, $\hat{M} \cup \{\{x_2, \hat{y}\}\}$ saturates all vertices in Y that are not on τ_1 or $\hat{M} \cup \{\{x_1, \hat{y}\}\}$ saturates all vertices in Y that are not on τ_3, a contradiction. Therefore, Cases 4 and 5 cannot occur.

Since these cases exhaust all possibilities, we have a contradiction. Therefore, no such counterexample G can exist, and the theorem is proved. □

4. Nonzero entries in $\mathcal{Q}$. In § 2, we have identified positions in $\mathcal{Q}$ that must be zero for a variety of more or less subtle reasons. To complete the solution to our problem and see that we have identified all the necessarily zero positions in $\mathcal{Q}$, we must show that for each remaining position there is a full rank matrix $A \in \mathcal{A}$ whose QR factorization has a nonzero entry in Q in the position of interest. Our strategy is, for *each* nonzero position, to identify a particular full rank matrix A which produces a Q with a nonzero in that position and whose pattern is *contained in* that of $\mathcal{A}$. A simple perturbation argument then yields an A with the *exact* pattern of $\mathcal{A}$. We note that our argument is sufficient to prove that the zero/nonzero structure for $\mathcal{Q}$ is tight (i.e., the best possible).

We first consider a particularly simple form for $\mathscr{A}$. To describe our construction, we introduce the notation Q^u for a matrix that has orthogonal columns that are *not* generally of unit length (i.e., unnormalized); Q_j^u denotes the jth column of this matrix.

LEMMA 4.1. *Let $\mathscr{A}$ be an $f \times f$ pattern with bipartite graph $B_f(\mathscr{A}) = (R_f(\mathscr{A}), C_f(\mathscr{A}); E_f(\mathscr{A}))$, in which $R_f(\mathscr{A}) = \{r_1, \ldots, r_f\}$, $C_f(\mathscr{A}) = \{c_1, \ldots, c_f\}$, and the edges in $E_f(\mathscr{A})$ form a simple path of length $2f-1$ from c_f to some r_i, $1 \leqq i \leqq f$. There exists a nonsingular matrix $A \in \mathscr{A}$ such that if $A = QR$, then q_{if} is nonzero.*

Proof. Assume that the path from c_f to r_i is

$$c_f = c_{j_f}, r_{k_f}, c_{j_{f-1}}, \ldots, c_{j_2}, r_{k_2}, c_{j_1}, r_{k_1} = r_i.$$

Let $A = [a_{de}]$ be the $f \times f$ (0, 1) matrix with $a_{de} = 1$ if and only if $\{r_d, c_e\}$ lies on the path; thus $A \in \mathscr{A}$. Matrix A has exactly two entries equal to 1 in each column (except the last). Since A is permutation equivalent to a lower triangular matrix with ones on the main diagonal, A is nonsingular. The columns (except c_f) can occur in any order on the path, but the exact order is important to the following construction of Q_f^u.

Consider

$$(4.1) \qquad Q_f^u = Q_{j_f}^u = fA_{j_f} - (f-1)A_{j_{f-1}} + \cdots \mp 2A_{j_2} \pm A_{j_1},$$

where the columns are from matrix A, the multiplying factors have alternating signs, and f is the number of columns on the path. Clearly, this $Q_{j_f}^u \in \operatorname{span}\{A_{j_1}, \ldots, A_{j_f}\}$, and we need to check the orthogonality conditions. By the construction of A, we have $A_{j_1} \cdot A_{j_1} = 2$, $A_{j_1} \cdot A_{j_2} = 1$, and $A_{j_1} \cdot A_{j_k} = 0$ for $k = 3, \ldots, f$. Thus $Q_f^u \cdot A_{j_1} = 0$. Similarly $A_{j_k} \cdot A_{j_k} = 2$ for $k = 2, \ldots, f-1$, and $A_{j_k} \cdot A_{j_{k-1}} = A_{j_k} \cdot A_{j_{k+1}} = 1$ for $k = 2, \ldots, f-1$, with A_{j_k} (combinatorially) orthogonal to all other columns. Thus $Q_f^u \cdot A_{j_k} = 0$ for $k = 1, \ldots, f-1$. The construction of A and the distinct multipliers in (4.1) mean that there can be no numerical cancellation, and so the ith entry of Q_f^u is nonzero. It is clear that if the column Q_f^u is normalized to have unit length, then column f of the matrix Q of the QR factorization of A is obtained, with $q_{if} \neq 0$. □

Note that the construction in Lemma 4.1 guarantees that all entries in column f of Q are nonzeros, not just q_{if}. To illustrate Lemma 4.1, consider an example in which $i = 2$ and $f = 3$.

Example 4.2. Consider the pattern $\mathscr{A}$ and bipartite graph $B_3(\mathscr{A})$ in Fig. 2. The edges in $E_3(\mathscr{A})$ form one path from column c_3 to r_2, namely, $c_3, r_3, c_1, r_1, c_2, r_2$. Take the matrix

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \in \mathscr{A}.$$

From (4.1), $Q_3^u = 3A_3 - 2A_1 + A_2 = (-1, 1, 1)^T$ (note that the order of the A_i in this equation is determined by the path), and the (2, 3) entry of Q^u is equal to 1. Thus, if $A = QR$, column 3 of Q must be a scalar multiple of Q_3^u (by the uniqueness of the QR factorization), and in fact

$$Q = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{bmatrix}.$$

$$\mathscr{A} = \begin{bmatrix} * & * & 0 \\ 0 & * & 0 \\ * & 0 & * \end{bmatrix} \quad \text{with } B_3(\mathscr{A}):$$

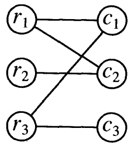


FIG. 2

Example 4.3. In the event that $\mathscr{A}$ has bipartite graph $B_f(\mathscr{A})$ with edges forming a path traversing the columns in order $c_f, c_{f-1}, \ldots, c_1$, then $A \in \mathscr{A}$ is bidiagonal. For example, when $f = 5$,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \in \mathscr{A},$$

the corresponding Q^u is the full upper Hessenberg matrix

$$Q^u = \begin{bmatrix} 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ 0 & 2 & 1 & -1 & 1 \\ 0 & 0 & 3 & 1 & -1 \\ 0 & 0 & 0 & 4 & 1 \end{bmatrix},$$

and note that R is bidiagonal upper triangular.

To prove our result for more general patterns, we need the fact that Q depends continuously on A. More specifically, if A is a matrix with full column rank and $A = QR$, then there exists a neighborhood of A such that every matrix in this neighborhood has full column rank and the orthogonal matrix of its factorization depends continuously on the entries. This qualitative fact (which is all we need) may be straightforwardly proven by elementary means, but precise bounds have been given in [S, Thm. 3.1]. We now use continuity to prove our main result on nonzero entries in $\mathscr{Q}$.

THEOREM 4.4. *For $m \geqq n$, let $\mathscr{A}$ be an $m \times n$ pattern with the Hall property. If $r_i \in p_j$, $1 \leqq j \leqq n$, then there is an $A \in \mathscr{A}$ of rank n such that $A = QR$ with q_{ij} nonzero.*

Proof. Without loss of generality, we can assume that each diagonal entry of $\mathscr{A}$ is $*$. (Thus, as noted in § 2, $s_j = \{i\colon 1 \leqq i \leqq j \text{ and } \mathscr{A}_i \in S_j\}$, and consequently, if some column $c_e \notin C_j(\mathscr{A})$, where $1 \leqq c_e < j$, then $r_e \notin R_j(\mathscr{A})$.) As $r_i \in p_j$, there exists at least one path in $B_j(\mathscr{A})$ from c_j to r_i. Thus, by Theorem 3.2, there exists a path, say

$$c_j = c_{j_f}, r_{k_f}, c_{j_{f-1}}, \ldots, c_{j_2}, r_{k_2}, c_{j_1}, r_{k_1} = r_i, \tag{4.2}$$

from c_j to r_i of length $2f - 1$ $(1 \leqq f \leqq j)$, and a matching M in $B_j(\mathscr{A})$ of all columns in $B_j(\mathscr{A})$ that are not on the path (4.2) to some subset of the rows in $B_j(\mathscr{A})$ that are not on the path (4.2). For any nonzero ε, let $A_\varepsilon = [a_{de}]$ be the $m \times j$ matrix whose pattern $\mathscr{A}_\varepsilon$ is the same as the first j columns of $\mathscr{A}$ and such that $a_{de} \in \{0, 1, \varepsilon\}$ with nonzero terms given by the following construction:

$$a_{de} = \begin{cases} 1 & \begin{array}{l} \text{if the edge } \{r_d, c_e\} \text{ is on the path (4.2),} \\ \textit{or} \text{ if the edge } \{r_d, c_e\} \in M, \\ \textit{or} \text{ if } d = e \text{ and } c_e \notin C_j(\mathcal{A}), \text{ the column vertex set of } B_j(\mathcal{A}); \end{array} \\ \varepsilon & \begin{array}{l} \text{if the } (d, e) \text{ entry of } \mathcal{A} \text{ is } * \text{ and } a_{de} \text{ has not been set} \\ \text{to 1 above.} \end{array} \end{cases}$$

(Note that $\mathcal{A}_{j+1}, \ldots, \mathcal{A}_n$ do not enter here.)

Let A_0 denote the matrix A_ε with all ε entries set to zero, and let $\tilde{A} = [\tilde{a}_{de}]$ be the (0, 1) square submatrix of A_0 restricted to rows $r_{k_1}, \ldots, r_{k_f}$ and columns $c_{j_1}, \ldots, c_{j_f}$, with $\tilde{a}_{de} = 1$ if and only if $a_{de} = 1$. Let $\tilde{\mathcal{A}}$ be the pattern of $\tilde{A}$; then $\tilde{\mathcal{A}}$ satisfies the conditions of Lemma 4.1 and $\tilde{A}$ is the matrix of the construction in Lemma 4.1. Thus, if $\tilde{A} = \tilde{Q}\tilde{R}$, then $\tilde{q}_{ij}$ is nonzero.

Now we consider the matrix A_0, and note that $\tilde{A}$ is embedded in A_0. By construction, columns of A_0 not on the path (4.2) contain distinct unit vectors, so A_0 has full column rank. If $A_0 = Q_0R_0$, then $\tilde{Q}$ is embedded in Q_0 in the same way as $\tilde{A}$ is embedded in A_0, with corresponding unit column vectors in columns not on the path. Thus the (i, j) entry of Q_0 is equal to $\tilde{q}_{ij}$ and is nonzero.

Finally, consider $A_\varepsilon = Q_\varepsilon R_\varepsilon$, which has the same column rank as A_0, namely j, for sufficiently small ε. By the continuity statement above, for sufficiently small ε, the matrix Q_ε has its (i, j) entry nonzero. Thus, any full rank matrix $A \in \mathcal{A}$ with the submatrix in its first j columns equal to A_ε has q_{ij} nonzero, where $A = QR$, since column j of Q is independent of $A_{j+1}, \ldots, A_n$. □

To illustrate this construction, consider the following example.

Example 4.5. Consider the pattern $\mathcal{A}$ and bipartite graph $B_4(\mathcal{A})$ in Fig. 3.

There is one path in $B_4(\mathcal{A})$ from c_4 to r_2, namely $c_4, r_4, c_1, r_1, c_2, r_2$, implying that the (2, 4) entry of $\mathcal{Q}$ is $*$. As this is the only path in $B_4(\mathcal{A})$ from c_4 to r_2, this must be the path (4.2) in the proof of Theorem 4.4, and the matching M can be chosen as either the edge $\{r_3, c_3\}$ or the edge $\{r_5, c_3\}$. If the first of these is chosen, then

$$A_\varepsilon = \begin{bmatrix} 1 & 1 & \varepsilon & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & \varepsilon & 0 \\ 0 & \varepsilon & 0 & 0 \end{bmatrix} \quad \text{and} \quad \tilde{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

$$\mathcal{A} = \begin{bmatrix} * & * & * & 0 & 0 \\ 0 & * & 0 & 0 & 0 \\ 0 & 0 & * & 0 & * \\ * & 0 & 0 & * & 0 \\ 0 & 0 & * & 0 & * \\ 0 & * & 0 & 0 & 0 \end{bmatrix}$$

with $B_4(\mathcal{A})$:

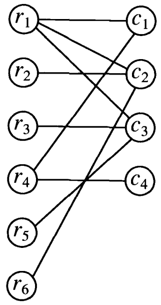


FIG. 3

From the proof of Lemma 4.1, the third column of $\tilde{Q}$ is a scalar multiple of $(-1, 1, 1)^T$, implying that the fourth column of Q_0 is a scalar multiple of $(-1, 1, 0, 1, 0, 0)^T$. Thus, by continuity, the $(2, 4)$ entry of Q_ε is nonzero, implying that the $(2, 4)$ entry of $\mathcal{Q}$ is nonzero.

Now consider a further example, with a pattern having more than one path from c_j to r_i (cf. Example 4.5 with a change in column 1).

Example 4.6. Consider the pattern $\mathcal{A}$ and bipartite graph $B_3(\mathcal{A})$ in Fig. 4. Focussing on the $(6, 3)$ entry, there are two paths, namely c_3, r_1, c_1, r_6 and c_3, r_1, c_2, r_6, from c_3 to r_6. The path (4.2) in the proof of Theorem 4.4 (i.e., the path τ in Theorem 3.2) is the first of these, and the matching M is the edge $\{r_2, c_2\}$. Note that for the other path from c_3 to r_6, there is no matching in the bipartite graph of c_1 with any row vertex not on this path (i.e., r_2, r_3, r_4, or r_5). This illustrates the care required in the selection of a path (4.2) in the proof of Theorem 4.4; there must exist an associated matching M so that A_ε is constructed such that A_0 has full column rank. The result of Theorem 3.2 ensures that this is always possible.

With the choice of the path c_3, r_1, c_1, r_6 above, we obtain

$$A_\varepsilon = \begin{bmatrix} 1 & \varepsilon & 1 \\ 0 & 1 & 0 \\ 0 & 0 & \varepsilon \\ 0 & 0 & 0 \\ 0 & 0 & \varepsilon \\ 1 & \varepsilon & 0 \end{bmatrix} \in \mathcal{A}.$$

For sufficiently small ε, if $A_\varepsilon = A = QR$, then q_{63} is nonzero.

We now combine the results of Theorems 2.1 and 4.4 to solve the problem stated in the introduction (as far as $\mathcal{Q}$ is concerned) in terms of the row index partition defined in § 2.

THEOREM 4.7. *For $m \geqq n$, let $\mathcal{A}$ be an $m \times n$ pattern with the Hall property. Then the (i, j) entry of $\mathcal{Q}$ is zero if and only if $i \in s_{j-1} \cup t_j \cup u_j$. Equivalently, the (i, j) entry of $\mathcal{Q}$ is nonzero if and only if $i \in p_j$.*

The next two corollaries indicate how $\mathcal{Q}$ can be computed. In the strong Hall case, $S_{j-1} = \phi$, $2 \leqq j \leqq n$, and $\mathcal{Q}_j$ is the union of $\mathcal{A}_j$ and certain columns $\mathcal{A}_w$, $1 \leqq w \leqq j - 1$.

COROLLARY 4.8. *For $m \geqq n$, let $\mathcal{A}$ be an $m \times n$ pattern with the strong Hall property. For any fixed j, $1 \leqq j \leqq n$, partition $\mathcal{A}_1, \ldots, \mathcal{A}_j$ into two disjoint sets $\mathcal{X}_j$ and $\mathcal{Y}_j$, for which $\mathcal{A}_j \in \mathcal{X}_j$, every column of $\mathcal{X}_j$ is combinatorially orthogonal to every column of $\mathcal{Y}_j$, and $|\mathcal{Y}_j|$ is as large as possible. Then $\mathcal{Q}_j = \cup\, \mathcal{A}_w$, $\mathcal{A}_w \in \mathcal{X}_j$.*

Proof. Since $S_{j-1} = \phi$, $2 \leqq j \leqq n$, the bipartite graph $B_j(\mathcal{A})$ has $C_j(\mathcal{A}) = \{c_1, \ldots, c_j\}$ and $R_j(\mathcal{A}) = \{r_i = i$: there exists a $*$ in row i in one of the columns of $C_j(\mathcal{A})\}$. For $k \leqq j$, column $c_k \in \mathcal{X}_j$ if and only if there exists a path from c_j to some

$$\mathcal{A} = \begin{bmatrix} * & * & * \\ 0 & * & 0 \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & * \\ * & * & 0 \end{bmatrix}$$

with $B_3(\mathcal{A})$:

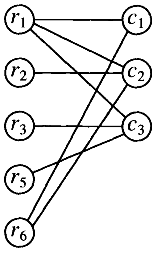


FIG. 4

r_i, where column c_k has a $*$ in row r_i. By definition of p_j, this is true if and only if $r_i \in p_j$; and by Theorem 4.7, $r_i \in p_j$ if and only if q_{ij} is nonzero. Together these give exactly the union statement above. □

For $\mathscr{A}$ having the Hall property, the essential modification to the result of the above corollary is that in the union of columns $\mathscr{A}_w$ specifying $\mathscr{Q}_j$, the column vectors in S_{j-1} and nonzeros in rows s_{j-1} must be omitted.

COROLLARY 4.9. *For $m \geqq n$, let $\mathscr{A}$ be an $m \times n$ pattern with the Hall property. For any fixed j, consider the subpattern of $\mathscr{A}$ in columns $\{\mathscr{A}_1, \dots, \mathscr{A}_j\} - S_{j-1}$ and rows $v_j = \{1, \dots, m\} - s_{j-1}$; this is an $(m - |s_{j-1}|) \times (j - |s_{j-1}|)$ pattern. Partition the columns of this subpattern into two disjoint sets $\mathscr{X}_j$ and $\mathscr{Y}_j$, with $\mathscr{A}_j[v_j] \in \mathscr{X}_j$, every column of $\mathscr{X}_j$ is combinatorially orthogonal to every column of $\mathscr{Y}_j$, and $|\mathscr{Y}_j|$ is as large as possible. Then $\mathscr{Q}_j[s_{j-1}] = 0$ and $\mathscr{Q}_j[v_j] = \cup \mathscr{A}_w[v_j]$, with $\mathscr{A}_w[v_j] \in \mathscr{X}_j$.*

Proof. By definition, p_j is the set of all row indices r_i such that

(i) $r_i \notin s_{j-1}$;

(ii) there exists a column c_k, $1 \leqq k \leqq j$, such that $\mathscr{A}_k \notin S_{j-1}$ and there is a $*$ in row r_i of c_k; and

(iii) there exists a path in $B_j(\mathscr{A})$ from c_j to r_i.

Thus the positions of nonzeros in vectors $\mathscr{A}_w[v_j] \in \mathscr{X}_j$ correspond precisely to entries in p_j, and the result follows from Theorem 4.7. □

The following algorithm implements the construction in Corollary 4.9.

ALGORITHM 4.10. Computation of the pattern $\mathscr{Q}$.

```
Notation:
    $\mathscr{A}$ is an $m \times n$ pattern, $m \geqq n$, satisfying the Hall property.
    $\mathscr{A}_j$ is the $j$th column of $\mathscr{A}$, $1 \leqq j \leqq n$.
    $\beta(\mathscr{A}_j) = \{i \mid$ the $i$th entry of $\mathscr{A}_j$ is $*\}$.
    $S_j$, $s_j$ are as defined in § 2.
    1. For $1 \leqq j \leqq n-1$, compute $S_j$ and $s_j$
    2. Initialization
           $\alpha_1 = \beta(\mathscr{A}_1)$
           $\beta(\mathscr{Q}_1) = \alpha_1$
           $\alpha_2 = \beta(\mathscr{A}_2) - s_1$
           If $\alpha_1 \cap \alpha_2 \neq \phi$, then $\alpha_2 = \alpha_2 \cup \alpha_1$
           $\beta(\mathscr{Q}_2) = \alpha_2$
    3. For $j = 3, 4, \dots, n$ do
       3.1. If $s_{j-1} \neq s_{j-2}$, then do
                For $k = 1, 2, \dots, j-1$ do
                    If $k \notin s_{j-1}$ then do
                        $\alpha_k = \beta(\mathscr{A}_k) - s_{j-1}$
                        For $i = 1, 2, \dots, k-1$ do
                            If $i \notin s_{j-1}$ and $\alpha_k \cap \alpha_i \neq \phi$ then $\alpha_k = \alpha_k \cup \alpha_i$
       3.2. $\alpha_j = \beta(\mathscr{A}_j) - s_{j-1}$
       3.3. For $i = 1, 2, \dots, j-1$ do
                If $i \notin s_{j-1}$ and $\alpha_j \cap \alpha_i \neq \phi$, then $\alpha_j = \alpha_j \cup \alpha_i$
       3.4. $\beta(\mathscr{Q}_j) = \alpha_j$
```

We illustrate this algorithm with the pattern $\mathscr{A}$ of Example 4.5. The only nonempty Hall set S_j, $1 \leqq j \leqq 4$, is $S_4 = \{\mathscr{A}_1, \mathscr{A}_4\}$, giving $s_4 = \{1, 4\}$. For $j = 4$, $s_2 = s_3 = \phi$, so computation of $\mathscr{Q}_4$ proceeds from step 3 of Algorithm 4.10 as follows. In step 3.2, $\alpha_4 = \beta(\mathscr{A}_4) = \{4\}$. In step 3.3, for $i = 1, 2, 3$, accumulate $\beta(\mathscr{A}_1), \beta(\mathscr{A}_2), \beta(\mathscr{A}_3)$ into α_4, giving $\alpha_4 = \{1, 2, 3, 4, 5, 6\}$, and so $\mathscr{Q}_4$ is full. For $j = 5$, as $s_4 \neq s_3$ in step 3.1, update $\alpha_2 = \{2, 6\}$ and $\alpha_3 = \{3, 5\}$. In step 3.2, $\alpha_5 = \beta(\mathscr{A}_5) - s_4 = \{3, 5\}$. This remains unchanged in step 3.3 because now $\alpha_5 \cap \alpha_2 = \phi$ and $\alpha_5 \cap \alpha_3 = \alpha_5$. Thus step 3.4 gives $\beta(\mathcal{Q}_5) = \alpha_5 = \{3, 5\}$. The entire pattern $\mathcal{Q}$ is given by

$$\mathcal{Q} = \begin{bmatrix} * & * & * & * & 0 \\ 0 & * & * & * & 0 \\ 0 & 0 & * & * & * \\ * & * & * & * & 0 \\ 0 & 0 & * & * & * \\ 0 & * & * & * & 0 \end{bmatrix}.$$

We now summarize a graph-theoretic way to compute the Hall sets required by step 1 of Algorithm 4.10. First, a maximal transversal in the $m \times n$ pattern $\mathcal{A}$ is obtained; see, e.g., [LP, p. 15] for an algorithm, or [D] for the case $m = n$. Without loss of generality, this maximum matching gives a square $n \times n$ pattern in which all diagonal entries are $*$. Second, for the digraph associated with this square pattern, find the strongly connected components using an algorithm of Tarjan (see, e.g., [RND]) and use these to construct its condensation digraph. The Hall sets of maximal cardinality can be determined from this condensation digraph by traversing paths starting at vertices of indegree zero.

The time complexity of this procedure to determine the Hall sets is dominated by the computation of a maximal transversal and is $O(\tau(m+n)^{1/2})$, where τ is the number of nonzero entries in $\mathcal{A}$ (see, e.g., [LP]). Step 3 of Algorithm 4.10 has time complexity $O((h+1)mn^2)$, where h is the number of distinct (nonempty) maximal Hall sets S_j, $1 \leqq j \leqq n-1$.

5. Pattern $\mathcal{R}$. Given a pattern $\mathcal{A}$, Theorem 4.7 determines the pattern $\mathcal{Q}$, and we now consider the pattern $\mathcal{R}$. As noted in the introduction, the pattern $\mathcal{R}$ is invariant under row permutations of $\mathcal{A}$.

THEOREM 5.1. *For $m \geqq n$, let $\mathcal{A}$ be an $m \times n$ pattern with the Hall property. If the pattern $\mathcal{Q}$ is determined as in Theorem* 4.7, *then the pattern $\mathcal{R}$ is given by the upper triangular part of $\mathcal{Q}^T\mathcal{A}$.*

Proof. Without loss of generality, we can assume that each diagonal position of $\mathcal{A}$ is $*$. Since $i \in p_i$, $1 \leqq i \leqq n$, the (i, i) entry of $\mathcal{Q}$ is $*$ by Theorem 4.7. Thus, since the (i, i) entry of $\mathcal{A}$ is $*$, the (i, i) entry of $\mathcal{Q}^T\mathcal{A}$ is $*$. The diagonal entries of $\mathcal{R}$ are all $*$ (by definition of the unique QR factorization), thus the patterns of $\mathcal{Q}^T\mathcal{A}$ and $\mathcal{R}$ agree on the diagonal.

Now consider in $\mathcal{R}$ a fixed (i, j) entry with $i < j \leqq n$. If $\mathcal{Q}_i$ is combinatorially orthogonal to $\mathcal{A}_j$, then $\mathcal{Q}_i^T\mathcal{A}_j = 0$. Also, for any $A \in \mathcal{A}$, if $A = QR$, then $R = Q^TA$ and so $r_{ij} = Q_i^TA_j = 0$ in this case of combinatorial orthogonality.

If $\mathcal{Q}_i$ is *not* combinatorially orthogonal to $\mathcal{A}_j$, then $\mathcal{Q}_i^T\mathcal{A}_j$ is $*$. We must show that there exists a matrix $A \in \mathcal{A}$ such that if $A = QR$, then $r_{ij} \neq 0$. Suppose that the (k, j) entry of $\mathcal{A}$ is $*$ for some k, $1 \leqq k \leqq m$, and the (k, i) entry of $\mathcal{Q}$ is $*$. Then $k \in p_i$, and by the construction in Theorem 4.4, the submatrix in the first i columns of A can be determined so that in its QR factorization, the entry $q_{ki} \neq 0$. (Note that in the QR factorization of any $A \in \mathcal{A}$, Q_i is completely determined by the first i columns of A.) Since $\mathcal{Q}_i$ and $\mathcal{A}_j$ are not combinatorially orthogonal, we may choose $A_j \in \mathcal{A}_j$ so that A_j is not orthogonal to the already determined Q_i. Thus, $r_{ij} = Q_i^TA_j$ is nonzero, completing the proof. (Note that only columns $A_1, \ldots, A_i$ and A_j have been specified; the other columns of A are arbitrary subject to $A \in \mathcal{A}$.) □

Returning again to Example 1.1, we see that Theorem 4.7 predicts the true pattern for $\mathcal{Q}$, that is, as in (1.2) with $\otimes = 0$. Theorem 5.1 then predicts the correct pattern for $\mathcal{R}$, again (1.2) with $\otimes = 0$. For example, the (4, 6) entry of $\mathcal{R}$ is zero, as $\mathcal{Q}_4$ is combinatorially orthogonal to $\mathcal{A}_6$.

The (i, j) entry of $\mathscr{R}$ is $*$ if and only if there is an index k such that $a_{kj} \neq 0$, and a path in $B_i(\mathscr{A})$ from column i to row k (passing only through columns numbered less than i). If $\mathscr{A}$ has the strong Hall property and $A \in \mathscr{A}$, this is equivalent (ignoring cancellation) to the existence of a path in the graph of A^TA from i to j passing only through vertices less than i. As the existence of such a path is equivalent to a generically nonzero entry in the (i, j) position of U in the LU factorization of A^TA, our result for $\mathscr{R}$, specialized to the strong Hall case, yields the same result as that of [CEG] mentioned in our introduction. We note that, since the columns of $\mathscr{Q}$ are determined sequentially (without appeal to previously determined columns), the rows of $\mathscr{R}$ may be determined serially. Thus, if only the pattern of $\mathscr{R}$ is of interest, it may be determined without necessity for storage of $\mathscr{Q}$. In fact, if only row i of $\mathscr{R}$ is desired, it may be determined from $\mathscr{Q}_i$ (and $\mathscr{A}$) for which only the Hall set S_{i-1} and $B_i(\mathscr{A})$ are necessary.

Remarks. We thank F. Ruskey, M. Gillespie, and the referees for useful comments. We also note that Lemma 2.5 of [Gi] (which was brought to our attention after preparation of this manuscript) is closely related in substance to our Theorem 3.2. The two proofs are quite different, and the underlying concept seems to be an important one for exhibiting nonzeros in factorizations of sparse matrices. The report [Gi] does not consider strong Hall critical matrices, as in our Theorem 3.1.

In addition, A. Pothen (private communication in response to our manuscript) proposed an alternate proof of our Theorem 4.4, based on the Dulmage–Mendelsohn (DM) decomposition and a statement equivalent to Lemma 2.5 of [Gi]. Pothen also notes that the DM decomposition can be used to construct the required Hall sets, and notes (via an argument based on algebraic indeterminates) that the collection of all entries that may individually be nonzero in Q may simultaneously be nonzero for some (in fact, almost all) allowed instances of A.

Note added in proof. After this paper was accepted for publication, we noticed that Problem 7.6 and its solution (in *Combinatorial Problems and Exercises*, L. Lovász, North-Holland, Amsterdam, 1979) could be applied (with $k = 1$) instead of our Theorem 3.1 in the proof of Theorem 3.2.

REFERENCES

[CEG] T. Coleman, A. Edenbrandt, and J. Gilbert, *Predicting fill for sparse orthogonal factorization*, J. Assoc. Comput. Mach., 33 (1986), pp. 517–532.

[D] I. S. Duff, *On algorithms for obtaining a maximum transversal*, ACM Trans. Math. Software, 7 (1981), pp. 315–330.

[Ge] W. M. Gentleman, *Row elimination for solving sparse linear systems and least squares problems*, in Proc. 1975 Dundee Conference on Numerical Analysis, G. A. Watson, ed., Lecture Notes in Math. 506, Springer-Verlag, New York, 1976, pp. 122–133.

[GH] A. George and M. Heath, *Solution of sparse linear least squares problems using Givens rotations*, Linear Algebra Appl., 34 (1980), pp. 69–83.

[GLN] A. George, J. Liu, and E. Ng, *Row-ordering schemes for sparse Givens transformations* I: *Bipartite graph model*, Linear Algebra Appl., 61 (1984), pp. 55–81.

[GN] A. George and E. Ng, *Symbolic factorization for sparse Gaussian elimination with partial pivoting*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 877–898.

[Gi] J. R. Gilbert, *An Efficient Parallel Sparse Partial Pivoting Algorithm*, Chr. Michelsen Institute Department of Science and Technology Report 88/45052-1, Bergen, Norway, 1988.

[HJ] R. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[JOD] C. R. Johnson, D. D. Olesky, and P. van den Driessche, *Inherited matrix entries*: *LU factorizations*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 94–104.

[LP] L. Lovász and M. D. Plummer, *Matching Theory*, North-Holland, Amsterdam, 1986.

[RND] E. M. Reingold, J. Nievergelt, and N. Deo, *Combinatorial Algorithms*: *Theory and Practice*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[S] G. W. Stewart, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 670–676, July 1993

005

ON SERIES EXPANSIONS AND STOCHASTIC MATRICES*

MOSHE HAVIV† AND Y. RITOV‡

Abstract. Let $P(0) \in R^{n \times n}$ be a stochastic matrix representing transition probabilities in a Markov chain, which is completely decomposable into m independent chains plus a number of transient states. Also, suppose that for all $\varepsilon > 0$ small enough $P(\varepsilon) \equiv P(0) + \varepsilon C$ is a stochastic matrix representing a unichain Markov process. Let $\pi(\varepsilon)$ be the stationary distribution of $P(\varepsilon)$ and let $Y(\varepsilon)$ be the deviation matrix of $P(\varepsilon)$ for $\varepsilon > 0$. It was proved by Schweitzer that $\pi(\varepsilon)$ has a series expansion around zero whose terms form a geometric sequence. He also showed that $Y(\varepsilon)$ admits a Laurent expansion. In order to compute the series expansion of $\pi(\varepsilon)$, a system of equations is defined resulting from equating coefficients of identical powers in the identity $\pi(\varepsilon)(I - P(\varepsilon)) = \underline{0}^T$. The authors prove that the minimal number of coefficients needed to be considered in order to get a system of equations that determines uniquely the leading term in the expansion for $\pi(\varepsilon)$ equals the order of the pole of $Y(\varepsilon)$ at zero plus one. Finally, the same system, but with a different right-hand side, determines the geometric factor of the series and hence the entire series expansion.

Key words. near uncoupling, series expansions, stochastic matrix

AMS subject classifications. 41, 60

1. Introduction. For ε, $0 < \varepsilon \leqq \varepsilon_{\max}$, let $P(\varepsilon) \equiv P(0) + \varepsilon C$ with $P(\varepsilon) \in R^{n \times n}$ be a *stochastic matrix*, representing transition probabilities in a unichain Markov process. Assume $P(0)$ to be stochastic but not necessarily representing a unichain Markov process. Let $\pi(\varepsilon) \in R^{1 \times n}$ be the *stationary distribution* of $P(\varepsilon)$ for $0 < \varepsilon \leqq \varepsilon_{\max}$. Thus $\pi(\varepsilon)$ uniquely satisfies

$$\pi(\varepsilon) = \pi(\varepsilon)P(\varepsilon)$$

and

$$\pi(\varepsilon)\underline{1} = 1,$$

where $\underline{1}$ is a vector of ones with the compatible dimension.

Also, let $Y(\varepsilon) \in R^{n \times n}$ be the *deviation matrix* of $P(\varepsilon)$ for $\varepsilon > 0$. Specifically, $Y(\varepsilon)$ is the unique matrix satisfying

$$Y(\varepsilon)(I - P(\varepsilon)) = I - E(\varepsilon) = (I - P(\varepsilon))Y(\varepsilon)$$

and

$$Y(\varepsilon)\underline{1} = \underline{0}, \qquad \pi(\varepsilon)Y(\varepsilon) = \underline{0}^T,$$

where $E(\varepsilon) = \underline{1}\pi(\varepsilon)$, namely, a matrix with all its rows equal to $\pi(\varepsilon)$, and where $\underline{0} \in R^{n \times 1}$ is a vector of zeros. We note here that for unichain $P(\varepsilon)$ the conditions $Y(\varepsilon)(I - P(\varepsilon)) = I - E(\varepsilon)$ and $Y(\varepsilon)\underline{1} = \underline{0}$ determine $Y(\varepsilon)$ uniquely once $E(\varepsilon)$ is given. The deviation matrix plays an important role in perturbation analysis of Markov chains and in the computation of mean passage times. See Schweitzer [17], Meyer [14], [15], or Haviv and Van der Heyden [10]. The matrix $Y(\varepsilon)$ is the group inverse of $I - P(\varepsilon)$. For more on group inverses, see Campbell and Meyer [1] .

Following a number of authors, especially Schweitzer [17], [18], [19] and Haviv and Ritov [9], we look for an issue involving the series expansion for $\pi(\varepsilon)$ and the

* Received by the editors February 11, 1991; accepted for publication (in revised form) December 12, 1991.

† Department of Econometrics, The University of Sydney, Sydney, NSW 2006 Australia (mosheh@sue.ucc.su.oz.au) and Department of Statistics, The Hebrew University, 91905 Jerusalem, Israel.

‡ Department of Statistics, The Hebrew University, 91905 Jerusalem, Israel (ritov@vms.huji.ac.il).

Laurent expansion for $Y(\varepsilon)$ for $\varepsilon > 0$ small enough. The case where $P(0)$ is unichain is relatively simple. From Schweitzer [17] we learn that for this case an expansion $\pi(\varepsilon) = \sum_{i=0}^{\infty} \pi^{(i)}\varepsilon^i$ for ε small enough exists with $\pi^{(0)}$ being $\pi(0)$, the stationary distribution of $P(0)$, and where for $i \geqq 1$, $\pi^{(i)} = \pi^{(i-1)}CY(0)$ with $Y(0)$ being the deviation matrix of $P(0)$. Finally, the deviation matrix $Y(\varepsilon)$ is analytic in some neighborhood of zero. In particular, $Y(\varepsilon) = \sum_{i=0}^{\infty} Y^{(i)}\varepsilon^i$ for $\varepsilon \geqq 0$ small enough for some matrices $Y^{(i)} \in R^{n\times n}$, $i \geqq 0$ with $Y^{(0)} = Y(0)$. Note that the single set of equations $\pi^{(0)}(I - P(0)) = \underline{0}^T$, resulting from equating the free coefficients in the identity $\pi(\varepsilon)(I - P(\varepsilon)) = \underline{0}^T$ plus the normalization condition $\pi(\varepsilon)\underline{1} = 1$, determine $\pi^{(0)}$ completely, and note that the order of the pole of $Y(\varepsilon)$ at zero is zero.

More involved is the case where $P(0)$ is not unichain, namely, $P(0)$ represents a Markov chain that is completely decomposable into a number of chains that do not interact one with the other, plus a number of transient states. In particular, note that in this case $\pi(0)$ is not well defined. We mention that in this case, $P(\varepsilon)$ for small values of ε is usually called in the literature *nearly uncoupled* or *nearly completely decomposable*. This model was looked at by a number of authors, for example, Courtois [4], Courtois and Louchard [5], Haviv and Van der Heyden [10], Vantilborgh [21], and for the question looked at here, mainly by Schweitzer [18], [19] and Haviv and Ritov [9]. See also Schweitzer [20], which surveys a large number of computational methods suitable for these models.

Next, we consider the case where $P(0)$ is completely decomposable. As noted in Schweitzer [18], the assumption that $P(\varepsilon)$ is unichain for $0 < \varepsilon \leqq \varepsilon_{\max}$ implies that $\pi(\varepsilon)$ is analytic in that region, and hence $\pi(\varepsilon)$ admits for that range a series expansion, $\sum_{i=0}^{\infty} \pi^{(i)}\varepsilon^i$, for some sequence $\{\pi^{(i)}\}_{i=0}^{\infty}$. Schweitzer also shows that $Y(\varepsilon) = \sum_{i=-n}^{\infty} Y^{(i)}\varepsilon^i$ for all $\varepsilon > 0$ small enough, namely, $Y(\varepsilon)$ has a pole at zero of order not greater than n.[1] The next question is how to compute the corresponding coefficients. Before commenting on that, we need to develop some notation. The complete decomposability assumption on $P(0)$ implies that it can be represented as

$$P(0) = \begin{pmatrix} Q_0 & Q_1 & Q_2 & \cdots & Q_m \\ 0 & P_1 & 0 & \cdots & 0 \\ 0 & 0 & P_2 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & P_m \end{pmatrix}$$

for some $m \geqq 2$, for some indecomposable matrices P_i, $1 \leqq i \leqq m$, and for some transient matrix Q_0. Next, we construct the eigensystems of $P(0)$ belonging to the eigenvalue 1. Let π_i be the unique vector satisfying $\pi_i = \pi_i P_i$ and $\pi_i\underline{1} = 1$. For $1 \leqq i \leqq m$, let $\vartheta_i \in R^{1\times n}$ be the vector $(\underline{0}^T, \ldots, \underline{0}^T, \pi_i, \underline{0}^T, \ldots, \underline{0}^T)$, where π_i appears at the indices corresponding to the independent chain i. Then let $V \in R^{m\times n}$ be a matrix whose ith row equals ϑ_i. Let $w_i^T \in R^{1\times n}$ be the vector $((I - Q_0)^{-1}Q_i\underline{1}^T, \underline{0}^T, \ldots, \underline{1}^T, \ldots, \underline{0}^T)$, where the vector $\underline{1}^T$ appears in the indices corresponding to the independent chain i. Finally, let $W \in R^{n\times m}$ be a matrix whose i column equals w_i. Note that $VP(0) = V$, $P(0)W = W$, $VW = I$, and $VP(0)W = I$, where $I \in R^{m\times m}$.

It is well known that both the algebraic and geometric multiplicities of the eigenvalue 1 of $P(0)$ equals the number of independent chains m, and hence V and W are full-rank matrices spanning the corresponding left and right eigenspaces. From the fact that $\pi^{(0)}(I - P(0)) = \underline{0}^T$, one gets that $\pi^{(0)} = z^{(0)}V$ for some $z^{(0)} \in R^{1\times m}$. Also, by equating

[1] Of course, if the order of the pole is k for some $k < n$, then $Y^{(i)} = 0$ for i, $-n \leqq i \leqq -(k+1)$.

the coefficients of ε in the identity $\pi(\varepsilon)(I-P(\varepsilon))=\underline{0}^T$, we get that $\pi^{(1)}(I-P(0))=\pi^{(0)}C$. From the analysis in Haviv and Ritov [9] or Haviv, Ritov, and Rothblum [11], we learn that these two facts imply that $z^{(0)}$ lies in the left null space of VCW. Hence, if this space is of dimension one, then we conclude that the first two sets of equations resulting from equating coefficients of identical powers are the minimal needed in order to determine $\pi^{(0)}$ uniquely. This case is analyzed by Schweitzer [18], who named this as the *full rank* case. Moreover, he showed that for this case $\pi^{(i)}=\pi^{(i-1)}U$ for $i\geqq 1$, where here $U\equiv CY[I-CW(VCW)^gV]$, and where $(VCW)^g$ is the group inverse of VCW.[2] Finally, Schweitzer also showed how to compute the Laurent expansion of the deviation matrix in the full rank case. In particular, he showed that $Y(\varepsilon)$ has a simple pole at zero. Again, as for the case when $P(0)$ is indecomposable, we get that the minimal number of coefficients needed to be considered in order to get a system of equations that determines $\pi(0)$ uniquely (up to a multiplicative factor, of course) equals the order of the pole of $Y(\varepsilon)$ at zero plus one.

Schweitzer [18] showed that if the Markov chains associated with $P(0)$ and $P(\varepsilon)$ for $0<\varepsilon\leqq\varepsilon_{\max}$ have the same set of transient states, then the full rank assumptions are met. However, the full rank assumptions do not hold in general. See § 3 for an example. Moreover, in many practical situations under the perturbation εC some states which are transient under $P(0)$ become recurrent and vice versa. See Rohlicek and Willsky [16] for some models in which this is typical. The aim of this paper is to establish the fact that is stated next as Theorem 1.1. Its proof is deferred until § 2.

THEOREM 1.1. *The minimal number of consecutive coefficients of various powers of ε in the identity $\pi(\varepsilon)(I-P(\varepsilon))=\underline{0}^T$ that need to be equated in order to get a system of equations that determines $\pi^{(0)}$ uniquely equals the order of the pole of $Y(\varepsilon)$ at zero plus one. Moreover, the same system of equations, but with a different right-hand side, determines the geometric factor matrix $U\in R^{n\times n}$ such that $\pi^{(i)}=\pi^{(0)}U^i$.*

Finally, we refer the reader to some other papers dealing with the same issue but from a different angle. Coderch, Willsky, Sastry, and Castanon [2], [3] and Rohlicek and Willsky [16] suggest a method for computing $\pi^{(0)}$. Another method for computing $\pi^{(0)}$ is given in Delebecque [6]. He extended it to the computation of the entire series $\{\pi^{(i)}\}_{i=0}^{\infty}$. The analysis of these authors who were concerned mainly with transient behavior of Markov processes is based on Kato's [12] classical perturbation results. We have already mentioned Schweitzer's work and note that an alternative method for computing the series expansion for $\pi(\varepsilon)$ and the Laurent expansion for $Y(\varepsilon)$ is suggested in Haviv and Ritov [9]. Also, in Hassin and Haviv [8] we can find a combinatorial algorithm, which finds the order of the poles of all the entries of $Y(\varepsilon)$ at zero. Finally, we mention Langenhop [13], who suggests a method for inverting near singular matrices. Specifically, note that for $\varepsilon>0$ small enough

$$\pi(\varepsilon)=(1,0,\dots,0)A(\varepsilon)^{-1}$$

and

$$Y(\varepsilon)=B(\varepsilon)A(\varepsilon)^{-1},$$

where $A(\varepsilon)$ is the matrix $I-P(\varepsilon)$ whose first column is replaced by a column of ones, and where $B(\varepsilon)$ is the matrix $E(\varepsilon)$ whose first row is replaced by zeros (Denardo [7]). Note that $A(0)$ is singular. Hence, we can invert $A(\varepsilon)$ by the method suggested by Langenhop and get the order of the pole of $Y(\varepsilon)$ as a byproduct.

[2] In Schweitzer [18] we can find a different expression. However, we omit a proof that both representations are equivalent.

2. Proof of Theorem 1.1. Solving for the unique $\pi(\varepsilon)=\sum_{i=0}^{\infty}\pi^{(i)}\varepsilon^{i}$, which satisfies $\pi(\varepsilon)=\pi(\varepsilon)P(\varepsilon)$ and $\pi(\varepsilon)\underline{1}=1$ for $\varepsilon>0$ small enough, is equivalent to finding the unique sequence $\{\pi^{(i)}\}_{i=0}^{\infty}$ with a positive radius of convergence, which satisfies

(0) $\pi^{(0)}(I-P(0))=\underline{0}^{T}$,

(1) $\pi^{(1)}(I-P(0))-\pi^{(0)}C=\underline{0}^{T}$,

$\vdots$

(i) $\pi^{(i)}(I-P(0))-\pi^{(i-1)}C=\underline{0}^{T}$,

$\vdots$

coupled with the normalization conditions $\pi^{(0)}\underline{1}=1$ and $\pi^{(i)}\underline{1}=0$ for $i\geqq 1$. The above systems of (difference) equations will be called the *fundamental equations*.

Schweitzer [18] proved that

$$Y(\varepsilon)=\sum_{i=-n}^{\infty}Y^{(i)}\varepsilon^{i}$$

for some sequence $\{Y^{(i)}\}_{i=-n}^{\infty}$. Specifically, as $Y(\varepsilon)$ equals $B(\varepsilon)A(\varepsilon)^{-1}$ (see the end of the introduction), each of its terms is a ratio between two polynomials, each of which has a degree not greater than n.

Let t be the minimal number of consecutive sets of fundamental equations needed to be considered in order to get a system of equations that determines $\pi^{(0)}$ uniquely.[3] This means that the set of fundamental systems of equations $(0,1,\ldots,t-1)$ plus the equations concerning the normalization requirements are sufficient in order to define a system of equations for which the corresponding solution for $\pi^{(0)}$ is unique. We first show that the order of the pole of $Y(\varepsilon)$ at zero is *at most* $t-1$, namely, that $Y^{(i)}=\underline{\underline{0}}$ for $-n\leqq i\leqq -t$, where $\underline{\underline{0}}\in R^{n\times n}$ is a matrix with all entries equal to zero.[4] Aiming for a contradiction, let j be the minimal index such that $Y^{(j)}\neq\underline{\underline{0}}$ for some $j\leqq -t$. Then consider the following set of t systems of equations, which result from equating the coefficients of $\varepsilon^{j},\varepsilon^{j+1},\ldots,\varepsilon^{j+t-1}$ in the matrix identity $Y(\varepsilon)(I-P(\varepsilon))=I-E(\varepsilon)$:

$$Y^{(j)}(I-P(0))=\underline{\underline{0}},$$
$$Y^{(j+1)}(I-P(0))-Y^{(j)}C=\underline{\underline{0}},$$
$$\vdots$$
$$Y^{(j+t-1)}(I-P(0))-Y^{(j+t-2)}C=\underline{\underline{0}}.$$

For example, look at the first row of each of the Y matrices here and consider the resulting equations. It is easy to see that these equations, coupled with the requirement that the sum of entries in that row be zero (resulting from the identity $Y(\varepsilon)\underline{1}=\underline{0}$), are like the set of equations for $\pi^{(0)},\pi^{(1)},\ldots,\pi^{(t-1)}$ but with a different right-hand side (in the normalization part), which is now identically zero. As the solution for $\pi^{(0)}$ is unique, so is the solution for $Y^{(j)}$. It is easy to see that $Y^{(j)}=Y^{(j+1)}=\cdots=Y^{(j+t-1)}=\underline{\underline{0}}$ solves this system, and hence $Y^{(j)}=\underline{\underline{0}}$ is the unique solution for $Y^{(j)}$.

Next we prove the converse inequality, namely, we show that the order of the pole of $Y(\varepsilon)$ at zero, which is denoted by k, is *at least* $t-1$. Let $\gamma^{(0)},\ldots,\gamma^{(k)}\in R^{1\times n}$ be *any* solution for the first $k+1$ sets of systems of fundamental equations, which also

[3] In Haviv and Ritov [9] it was proved that this number is bounded by $n+1$.

[4] This inequality is stated in Schweitzer [19].

satisfies $\gamma^{(0)}\underline{1} = 1$. Of course, for some sequence $\{\delta^{(i)}\}_{i=0}^{\infty}$ in $R^{1\times n}$ with $\delta^{(0)}\underline{1} = 0$, for all $\varepsilon > 0$ small enough, $\pi(\varepsilon) = \gamma(\varepsilon) + \delta(\varepsilon)$ where $\gamma(\varepsilon) = \sum_{i=0}^{k} \gamma^{(i)}\varepsilon^i$ and where $\delta(\varepsilon) = \sum_{i=0}^{\infty} \delta^{(i)}\varepsilon^i$. We will complete our proof by showing that $\delta^{(0)} = \underline{0}^T$. First, note that

$$[\gamma(\varepsilon) + \delta(\varepsilon)][I - P(0) - \varepsilon C] = \underline{0}^T.$$

Since $\{\gamma^{(i)}\}_{i=0}^{k}$ solves the system of equations resulting from the first $k+1$ sets of fundamental equations, the above identity implies that

$$\delta(\varepsilon)(I - P(0) - \varepsilon C) = \gamma^{(k)} C \varepsilon^{k+1},$$

and then that

$$\delta(\varepsilon) = \gamma^{(k)} C Y(\varepsilon) \varepsilon^{k+1} + \alpha(\varepsilon)\pi(\varepsilon)$$

for some $\alpha(\varepsilon) = \sum_{i=0}^{\infty} \alpha^{(i)}\varepsilon^i \in R$.[5] But $\delta^{(0)}\underline{1} = 0$ and $Y(\varepsilon)\underline{1} = \underline{0}$ and since $\pi^{(0)} \neq \underline{0}^T$ we conclude that $\alpha^{(0)} = 0$. As the order of the pole of $Y(\varepsilon)$ is k, the above right-hand side is of the order of magnitude of $\mathrm{O}(\varepsilon)$ and hence $\delta^{(0)} = \underline{0}^T$ as required.

We conclude the proof by establishing the fact that the geometric matrix factor can be derived from the same set of equations. Suppose we have determined the system of equations that determines $\pi^{(0)}$ uniquely and suppose we compute $\pi^{(1)}$. By considering the fundamental sets $(1, 2, \ldots, t)$, introducing the normalization requirements $\pi^{(i)}\underline{1} = 0$ for $1 \leqq i \leqq t$ and treating $\pi^{(0)}$ as known, we get the same system as the one considered previously when solving for $\pi^{(0)}$ with a different right-hand side. Since the solution for $\pi^{(0)}$ is unique, the same is now the case for $\pi^{(1)}$. In particular, if $\pi^{(0)}$ was computed by some matrix inversion, the same matrix inverse can be used in order to compute $\pi^{(1)}$. Also, $\pi^{(1)}$ is a linear function of $\pi^{(0)}$, and hence $\pi^{(1)} = \pi^{(0)}U$ for some matrix $U \in R^{n\times n}$.[6] By considering the fundamental sets $(2, 3, \ldots, t+1)$ and the normalization requirements, we get that $\pi^{(1)}$ plays the same role in the computation of $\pi^{(2)}$ previously played by $\pi^{(0)}$ in the computation of $\pi^{(1)}$. In particular, $\pi^{(2)} = \pi^{(1)}U$ for the same matrix U. Likewise, for all $i \geqq 1$, $\pi^{(i)} = \pi^{(i-1)}U$ or $\pi^{(i)} = \pi^{(0)}U^i$.

3. A numerical example. The following example is taken from Haviv and Ritov [9]. More details are given there. This example is also considered in Hassin and Haviv [8]. Let

$$P(\varepsilon) = P(0) + \varepsilon C = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \varepsilon \begin{pmatrix} 0 & -1 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Here

$$V = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

and hence

$$VCW = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

[5] The last observation is based on the fact that if A^g is the group inverse of A and the system of linear equations $xA = b$ is feasible, then x is a solution if and only if $x = bA^g + y$ for some vector y in the left null space of A. See, e.g., Campbell and Meyer [1].

[6] For an explicit expression for U, see Haviv and Ritov [9].

We can see that the full rank assumptions are not met, as the null space of VCW is of dimension larger than one. Note also that $P(0)$ and $P(\varepsilon)$ for $\varepsilon > 0$ do not possess the same set of transient states.

It is easy to check that the two conditions $\pi^{(0)}(I-P(0)) = \underline{0}^T$ and $\pi^{(1)}(I-P(0)) - \pi^{(0)}C = \underline{0}^T$ do not determine $\pi^{(0)}$ uniquely. Specifically, for any scalars α, β, γ, and δ the vectors $\pi^{(0)} = (0, \alpha, 0, \beta)$ and $\pi^{(1)} = (\alpha, \gamma, \beta, \delta)$ are feasible. Then, introducing the third system, $\pi^{(2)}(I - P(0)) - \pi^{(1)}C = \underline{0}^T$, restricts the solution to $\pi^{(0)} = (0, \alpha, 0, \alpha)$ for any α. Finally, the normalization requirement $\pi^{(0)}\underline{1} = 1$ leads to the unique solution $\pi^{(0)} = (0, \frac{1}{2}, 0, \frac{1}{2})$. The matrix U equals

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix},$$

and since $U^i = (-1)^{i+1}U$, we get that $\pi^i = (-1)^{i+1}(.5, -.5, .5, -.5)$.

By Theorem 1.1 the pole of $Y(\varepsilon)$ at zero is 2. Indeed, as computed in Haviv and Ritov [9],

$$Y^{(-2)} = \begin{pmatrix} 0 & .25 & 0 & -.25 \\ 0 & .25 & 0 & -.25 \\ 0 & -.25 & 0 & .25 \\ 0 & -.25 & 0 & .25 \end{pmatrix}.$$

REFERENCES

[1] S. L. CAMPBELL AND C. D. MEYER, *Generalized Inverses of Linear Transformations*, Pitman, London, 1979.

[2] M. CODERCH, A. S. WILLSKY, S. S. SASTRY, AND D. A. CASTANON, *Hierarchical aggregation of singularity perturbed finite state Markov processes*, Stochastics, 8 (1983), pp. 259–289.

[3] ———, *Hierarchical aggregation of linear systems with multiple time scale*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 1017–1030.

[4] P. J. COURTOIS, *Decomposability*, Academic Press, New York, 1977.

[5] P. J. COURTOIS AND G. LOUCHARD, *Approximation of eigencharacteristics on nearly decomposable systems*, Stochastic Process. Appl., 4 (1976), pp. 283–296.

[6] F. DELEBECQUE, *A reduction process for perturbed Markov chains*, SIAM J. Appl. Math., 43 (1983), pp. 325–350.

[7] E. V. DENARDO, *A Markov decision problem*, in Mathematical Programming, T. C. Hu and S. M. Robinson, eds., Academic Press, New York, 1973.

[8] R. HASSIN AND M. HAVIV, *Mean passage times and nearly uncoupled Markov chains*, SIAM J. Discrete Math., 5 (1992), pp. 386–397.

[9] M. HAVIV AND Y. RITOV, *Series Expansions for Stochastic Matrices*, unpublished manuscript.

[10] M. HAVIV AND L. VAN DER HEYDEN, *Perturbation bounds for the stationary probabilities of a finite Markov chain*, Adv. in Appl. Probab., 16 (1984), pp. 804–818.

[11] M. HAVIV, Y. RITOV, AND U. G. ROTHBLUM, *Taylor expansions of eigenvalues of perturbed matrices with applications to spectral radii of nonnegative matrices*, Linear Algebra Appl., 168 (1992), pp. 159–188.

[12] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1966.

[13] C. E. LANGENHOP, *The Laurent expansion for a nearly singular matrix*, Linear Algebra Appl., 4 (1971), pp. 329–340.

[14] C. D. MEYER, *The role of the group generalized inverse in the theory of finite Markov chains*, SIAM Rev., 17 (1975), pp. 443–464.

[15] ———, *The solution of a finite Markov chain and perturbation bounds for the limiting probabilities*, SIAM J. Algebraic Discrete Meth., 1 (1980), pp. 273–283.

[16] J. R. ROHLICEK AND A. S. WILLSKY, *The reduction of Markov generators: An algorithm exposing the role of transient states*, J. Assoc. Comput. Mach., 35 (1988), pp. 675–696.

[17] P. J. SCHWEITZER, *Perturbation theory and finite Markov chains*, J. Appl. Probab., 5 (1968), pp. 401–413.

[18] ———, *Perturbation series expansions of nearly completely decomposable Markov chains*, Working Papers Series No. 8122, The Graduate School of Management, The University of Rochester, New York, 1981.

[19] ———, *Perturbation series expansions of nearly completely decomposable Markov chains*, in Teletraffic Analysis and Computer Performance Evaluation, O. J. Boxma, J. W. Cohen, and H. C. Tijm, eds., Elsevier Science Publishers B.V., North-Holland, Amsterdam, 1986.

[20] ———, *A survey of aggregation/disaggregation in large Markov chains*, in Numerical Solution of Markov Chain, W. J. Stewart, ed., Marcel Dekker, New York, 1991, pp. 63–88.

[21] H. VANTILBORGH, *Aggregation with an error of* $O(\varepsilon^2)$, J. Assoc. Comput. Mach., 32 (1985), pp. 161–190.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 677-687, July 1993

THE NUMERICAL EFFECT OF MEASUREMENT ERROR IN THE EXPLANATORY VARIABLES ON THE OBSERVED LEAST SQUARES ESTIMATE*

SAMPRIT CHATTERJEE† AND GLENN HELLER‡

Abstract. The numerical effect of measurement error on the least squares estimate in the linear regression model is examined. The change in the least squares estimate is measured by calculating a stochastic upper bound on the relative distance between the true (unobserved) and observed (with error) jth component. The bound is derived in the case of k of p explanatory variables measured with error. If the bound indicates that none of the estimates are badly perturbed, the analysis can continue without concern about the effect of measurement error. Simulations are carried out to compare this bound with the first-order upper bound of Golub and Van Loan [*Matrix Computations*, Johns Hopkins University Press, 1983], and the componentwise upper bound of Higham [*Contemp. Math.*, 112 (1990), pp. 195-208].

Key words. measurement error, upper bound, perturbation theory, stochastic ordering, linear regression

AMS subject classifications. 62J05, 15A60, 15A45

1. Introduction and notation.

Throughout this paper we use upper case letters for matrices, lower case letters for scalars, and lower case bold letters for vectors. For a real matrix $A = (a_{ij})$ with n rows and m columns, the following definitions and inequalities involving matrix norms are used [10]:

$$\|A\|_2 = \sup_{\|\mathbf{v}\|=1} \|A\mathbf{v}\|_2, \tag{1a}$$

$$\|A\|_\infty = \max_i \sum_j |a_{ij}|, \tag{1b}$$

$$\|A\|_F^2 = \sum_i \sum_j a_{ij}^2, \tag{1c}$$

$$\|A\|_2 \leqq \sqrt{n}\|A\|_\infty, \tag{1d}$$

$$\|A\|_2 \leqq \|A\|_F. \tag{1e}$$

In addition, we denote the expectation and variance of a random variable by $E(\cdot)$ and var $(\cdot)$, respectively.

Given the classical linear regression model $y_i = \mathbf{z}_i^T\beta + \varepsilon_i$ $(i = 1, \dots, n)$, the least squares estimator of the regression coefficient vector β is $\mathbf{b}^z = (Z^TZ)^{-1}Z^Ty$, $Z^T = (\mathbf{z}_1, \dots, \mathbf{z}_n)$. Here $\mathbf{z}_i^T = (z_{i0}, z_{i1}, \dots, z_{ip})$ is a row vector denoting the ith observation of the p explanatory variables with $z_{i0} = 1$, and the $\{\varepsilon_i\}$ are independent identically distributed random variables with mean zero and variance σ^2. Often the variables $\mathbf{z}_i$ in the regression model are contaminated with measurement error, so the observable variables $\mathbf{x}_i = \mathbf{z}_i + \mathbf{u}_i$ are utilized. As a result, the observed least squares estimator is $\mathbf{b}^x = (X^TX)^{-1}X^Ty$, where $X^T = (\mathbf{x}_1, \dots, \mathbf{x}_n)$.

There is a substantial body of literature devoted to the development of measurement error models for linear regression, and many estimators have been proposed as an alter-

* Received by the editors January 14, 1991; accepted for publication (in revised form) December 17, 1991.

† Department of Statistics and Operations Research, New York University, 90 Trinity Place, New York, New York 10006 (schatterjee@stern.nyu.edu).

‡ Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York 10021 (heller@biost.mskcc.org).

native to the observed least squares estimator $b^{\mathbf{x}}$. Fuller [7] gives an overview of these models. Often, however, the assumptions needed to implement these measurement error models are untenable. A typical assumption is that the covariance matrix of the independent identically distributed rows of the error matrix $(\varepsilon \,|\, U)$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$, $U^T = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$, is known up to a scalar. If such an assumption cannot be made, then the analyst will be led to use the well-defined observed least squares estimate $\mathbf{b}^{\mathbf{x}}$. However, is $\mathbf{b}^{\mathbf{x}}$ a reasonable estimate to use when measurement error is present?

One approach to answering this question is explored in a series of papers by Gallo [8], Carroll, Gallo, and Gleser [2], and Gleser, Carroll, and Gallo [9]. These papers provide necessary and sufficient conditions for the linear combination $\mathbf{c}^{\mathbf{T}}\mathbf{b}^{\mathbf{x}}$ to be a consistent estimator of $\mathbf{c}^{\mathbf{T}}\beta$, and the condition when such linear combinations are asymptotically normally distributed. They show that the class of linear combinations is restricted, and they also note that $\mathbf{c}$ is a vector function of the limiting value of $n^{-1}X^TX$ and may not be known. A special case of their work proves the well-documented result that the jth observed coefficient b_j^x is not a consistent estimator of β_j.

A second approach, often called small sample asymptotics, or kth-order perturbation theory, assumes that $s\|U\|_2^{k+1}$ is negligible for fixed n and scalar s. The motivation for this method is that if the elements of U are small relative to the smallest eigenvalue of X^TX matrix, then the unobserved Z^TZ matrix is nonsingular and $\mathbf{b}^{\mathbf{z}}$ is well defined. An advantage to this methodology is that $(Z^TZ)^{-1}$ may be expanded in a series up to the kth-order term allowing for algebraic manipulation inside the inverse function. A first-order upper bound on the distance between the jth elements of the vectors $\mathbf{b}^{\mathbf{x}}$ and $\mathbf{b}^{\mathbf{z}}$ based on the 2-norm is [10]

$$(2) \qquad |\mathbf{b}_j^x - \mathbf{b}_j^z| \leqq \|\mathbf{e}_j^T(X^TX)^{-1}X^T\|_2\|U\|_2\|\mathbf{b}^x\|_2 + \|\mathbf{e}_j^T(X^TX)^{-1}\|_2\|U\|_2\|\mathbf{y} - X\mathbf{b}^x\|_2,$$

where $\mathbf{e}_j$ is the jth unit basis vector. This inequality is obtained by setting $\varepsilon = \|U\|_2/\|X\|_2$ in [10, p. 143]. Other examples of the benefits of this expansion are found in Stewart [15], [17] and Higham and Stewart [12], who derive a first-order upper bound for the case of only one explanatory variable measured with error. (Stewart [16] drops the first-order assumption and examines the problem of measurement error in the linear functional model, i.e., $\varepsilon = 0$. However, he still maintains the restriction that only one explanatory variable is measured with error.) A second-order perturbation expansion was computed by Hodges and Moore [13] to examine a bound on the bias of the observed least squares estimator $E(\mathbf{b}^{\mathbf{x}} - \mathbf{b}^{\mathbf{z}})$. However, their result gives the bias in terms of $(Z^TZ)^{-1}$ and β, which are unknown.

Two related approaches are developed by Davies and Hutton [5] and Fletcher [6]. Davies and Hutton [5] calculate a bound on the asymptotic bias of $\mathbf{b}^{\mathbf{x}}$, $\lim_{n\to\infty} E(\mathbf{b}^{\mathbf{x}} - \mathbf{b}^{\mathbf{z}})$. The large sample asymptotics used to bound the asymptotic bias provide the same benefit as the second-order calculations of Hodges and Moore [13]. Fletcher [6] provides a first-order upper bound on the mean squared error function $E\|\mathbf{b}^x - \mathbf{b}^z\|_2^2$. The first-order bound is based on the assumption that the measurement errors u_{ij} are independently distributed with mean zero and known variance.

A third approach examines componentwise perturbations in the covariate matrix. Higham [11] assumes that the measurement errors u_{ij} obey bounds of the form $|u_{ij}| < \delta a_{ij}$. The bound, which is developed using primarily triangle inequalities, is

$$(3) \qquad |\mathbf{b}_j^x - \mathbf{b}_j^z| \leqq \delta\{|\mathbf{e}_j^T(X^TX)^{-1}X^T|A|\mathbf{b}^z| + |e_j^T(X^TX)^{-1}|A^T|\mathbf{y} - Z\mathbf{b}^z|\},$$

where $|\mathbf{x}| = (|x_i|)$, and $a_{ij} = a_j$ for each i. Although the upper bound given in (3) is rigorous, it is a function of the unobserved regression coefficient $\mathbf{b}^z$ and residual $\mathbf{r}^z =$

$\mathbf{y} - Z\mathbf{b}^z$. Higham's suggestion is to estimate $\mathbf{b}^z$ and $\mathbf{r}^z$ with $\mathbf{b}^x$ and $\mathbf{r}^x$. Note that the bound is no longer rigorous after this estimation.

In this paper we have chosen an alternative approach that combines elements of the previous methods to answer the question, "Is $\mathbf{b}^x$ a reasonable estimate when measurement error is present?" A stochastic upper bound on a distance measure between the true and observed least squares estimate of β_j, in the general case of k explanatory variables measured with error, is derived. This bound does not rely on first-order expansions, nor does it require that the sample be large. When the upper bound is small, measurement error can be ignored and β_j is well estimated by b_j^x. From here on, unless otherwise specified, all variables are centered.

2. An upper bound for the relative error. A useful measure of the distance between b_j^x and b_j^z is the relative error

$$\frac{|b_j^x - b_j^z|}{|b_j^x|}.$$

However, since the true covariate matrix Z is unknown, the relative error cannot be computed. Instead, an upper bound of this metric is used to gauge the effect of measurement error on the least squares coefficients. Theorem 1 provides an upper bound for the relative distance between b_j^x and b_j^z in the general case of k explanatory variables measured with error. The proof of this theorem is given in the Appendix.

THEOREM 1. *Let* $d = \|X^+\|_2 \|U\|_2$. *If* $d < 1$, *and*

$$\|\mathbf{e}_j^T X^+ P_z\|_2^2 \leqq \|\mathbf{e}_j^{\mathrm{T}} Z^+\|_2^2 \leqq \|\mathbf{e}_j^{\mathrm{T}} X^+\|_2^2 [1 + \{\|\mathbf{e}_j^T (X^+ - Z^+)\|_2^2 / \|\mathbf{e}_j^T X^+\|_2^2\}],$$

then

$$\frac{|b_j^x - b_j^z|}{|b_j^x|} \leqq \Gamma_j \frac{d[\{d(2-d)\}^2 + (1-d)^2]^{1/2}}{|\cos \gamma_j|(1-d)}, \tag{4}$$

where $X^+ = (X^T X)^{-1} X^T$ *(*X *is assumed to be full column rank),* $P_z = ZZ^+$ *is a projection matrix,* γ_j *is defined as the angle between* $(X^+)^T \mathbf{e}_j$ *and* $\mathbf{y}$, *and* Γ_j *is defined by* $\cos\{\cos^{-1}(d) - \gamma_j\}$.

The cosine and Γ functions can be interpreted as correlation measures. It can be shown that the sample correlation r_j, of $\mathbf{y}$ and $(X^+)^T \mathbf{e}_j$, the component of $\mathbf{x}_\mathrm{j}$ orthogonal to $X_{[j]}$, (the $nx(p-1)$ matrix formed from the observed (centered) covariate matrix X by removing its jth column) is equal to $\cos \gamma_j$. In addition, if it is assumed that d is small, then $\Gamma_j \approx (1 - r_j^2)^{1/2}$.

Although $\|X^+\|_2$, $\cos \gamma_j$, and Γ_j can be calculated from the observed data, the measurement error matrix U is not observable and, consequently, d is not calculable. However, if an assumption is made on the precision of each of the p measurement instruments, further progress can be made. We assume that

$$|u_{ij}^R| < a_j \quad (i = 1, 2, \ldots, n) \quad (j = 1, 2, \ldots, p) \tag{5}$$

(f^R represents an uncentered variable). This assumption implies that it is possible to specify for each variable an upper bound on the absolute value of the measurement error. This is a realistic assumption after gross errors have been eliminated. Then from (1d),

$$\|U\|_2 \leqq \sqrt{n}(\textstyle\sum a_j),$$

and the bound for $\|U\|_2$ can be used in (4) to derive a calculable upper bound for the relative error.

Unfortunately, this bound is conservative in that it assumes that, for each observation, the maximum possible error occurs in each of the variables simultaneously. This is an event that occurs with very low probability. The approach taken in § 3 to tighten this bound is to assume that the errors are stochastic, and to consider bounds that hold $100(1-\delta)\%$ of the time. This enables the analyst to ignore large values of u_{ij} found in the upper tail of the measurement error distribution. We now derive a bound for $\|U\|_2$ under weak probability assumptions about measurement errors.

3. A stochastic upper bound. The bound on $\|U\|_2$ can be sharpened if we make some assumptions about the probability law of the measurement errors, and allow the bound to hold with probability $1-\delta$. To compute the bounds we need the following definition and lemma.

DEFINITION. The random variable w is stochastically larger than the random variable v if

$$\Pr[w>c] \geqq \Pr[v>c] \quad \text{for all } c.$$

For a discussion on stochastic ordering, see [14].

LEMMA 1. *Let the random variable w be uniformly distributed with support in $[-a,a]$. Among the class of random variables v with probability density functions symmetric about zero, nonincreasing in $|v|$, and support $[-a,a]$, w^2 is stochastically larger than v^2.*

The proof of Lemma 1 is given in the Appendix.

Assuming that the measurement errors u_{ij} are independent random variables ($i=1,2,\ldots,n;j=1,2,\ldots,p$) satisfying Lemma 1; then from (1e), we can find a constant c such that

$$\Pr[\|U\|_2^2<c] \geqq \Pr[\|W\|_F^2<c]=1-\delta,$$

where $w_{ij}\sim \mathrm{Unif}[-a_j,a_j]$ are independent. Thus $\|U\|_2^2$ is bounded by c with probability greater than $1-\delta$. Although c can be found explicitly, the calculations needed are difficult. Utilizing the first two moments of w_{ij}^2,

$$E(w_{ij}^2)=\frac{a_j^2}{3}, \qquad \mathrm{var}(w_{ij}^2)=\frac{4a_j^4}{45},$$

and the central limit theorem (assuming the Lindeberg–Feller conditions), a simple approximation to c is

$$c=(Z_{1-\delta})\left[\frac{4n}{45}\sum a_j^4\right]^{1/2}+\frac{n}{3}\sum a_j^2, \tag{6}$$

where $Z_{1-\delta}$ is the $(1-\delta)$ quantile of the standard normal distribution.

By replacing c for $\|U\|_2^2$ in the upper bound of Theorem 1, we have exchanged a rigorous bound for a probabilistic bound. However, if δ is kept small, the analyst can be confident that the relative error will not exceed this upper bound with a very high probability. Simulations executed in § 5 demonstrate this point.

COROLLARY 1. *Let $q^2=c\|X^+\|_2^2$. If $q^2<1$, and*

$$\|\mathbf{e}_j^T X^+ P_z\|_2^2 \leqq \|\mathbf{e}_j^T Z^+\|_2^2 \leqq \|\mathbf{e}_j^T X^+\|_2^2[1+\{\|\mathbf{e}_j^T(X^+-Z^+)\|_2^2/\|\mathbf{e}_j^T X^+\|_2^2\}],$$

then

$$\frac{|b_j^x-b_j^z|}{|b_j^x|} \leqq \Gamma_j\frac{q[\{q(2-q)\}^2+(1-q)^2]^{1/2}}{|\cos\gamma_j|(1-q)} \tag{7}$$

with probability $1-\delta$.

The relative error is invariant to column scaling of the carrier matrices X, Z and hence provides freedom to rescale for the purpose of reducing the upper bound in (7). In particular, only q^2 is affected by the choice of the column scaling matrix $D = \operatorname{diag}\{d_j\}$. Let

$$c_D = (Z_{1-\delta})\left[\frac{4n}{45}\sum (d_j a_j)^2\right]^{1/2} + \frac{n}{3}\sum (d_j a_j)^2,$$

$$\|(XD)^+\|_2^2 = \|D^{-1}X^+\|_2^2.$$

Although an optimal scaling matrix is unknown, one strategy is to choose a contraction mapping ($0 < d_j \leqq 1; j = 1, 2, \dots, p$), which reduces c_D but maintains $\|(XD)^+\|_2$ in the neighborhood of $\|X^+\|_2$ [18]. The theorem given below and proved in the Appendix specifies $\tilde{D}$ so that $\|(X\tilde{D})^+\|_2$ is never farther than $\sqrt{p}$ from its lower bound.

THEOREM 2. *Let $\phi(X^+) = \max_j \|\mathbf{e}_j^T X^+\|_2$. Choose the diagonal matrix $\tilde{D}$, so that the rows of $\tilde{D}^{-1}X^+$ are of equal length, and $\tilde{d}_j = 1$ for some j. Then*

$$\text{(i)}\quad \phi(\tilde{D}^{-1}X^+) = \inf_D \{\|D^{-1}X^+\|_2 : 0 < d_j \leqq 1\},$$

$$\text{(ii)}\quad \phi(\tilde{D}^{-1}X^+) \leqq \|\tilde{D}^{-1}X^+\|_2 \leqq \sqrt{p}\,\phi(\tilde{D}^{-1}X^+).$$

Remark. The equilibration matrix $\tilde{D}$ is generally defined up to a multiplicative constant. The constraint $\tilde{d}_j = 1$ for some j provides a unique $\tilde{D}$.

4. An illustrative example. We illustrate the numerical effect of measurement error on the least squares estimates using the abrasion loss data (see [3]). This data set examines the relationship between the response variable (y) abrasion loss (in grams per horsepower-hour) and the explanatory variables, (x_1) hardness (in degrees Shore) and (x_2) tensile strength (in kg/sq. cm). The unit-degrees Shore represents the angle a Shore hammer reaches at its maximum after bouncing off a rubber surface. Both variables are significant predictors of abrasion loss, and the uncentered regression equation is

$$\hat{y}_i = 885 - 6.57x_{i1}^R - 1.37x_{i2}^R.$$

Suppose that the instrument measuring the angle of the Shore hammer was accurate to within ±.5 units, and the instrument measuring the tensile strength was accurate to within ±1 unit. Then, letting $\delta = .05$ in (7), there is a probability of .95 that the upper bounds for the relative distance of b_j^x and b_j^z ($j = 1, 2$) are ($q = .047$)

$$\frac{|b_j^x - b_j^z|}{|b_j^x|} \leqq \begin{cases} .041, & j = 1, \\ .107, & j = 2. \end{cases}$$

Hence the regression coefficients are measured correctly to within ±.25.

5. Simulations. We conducted a series of simulations to compare our estimate of the maximum relative error (7) (after rescaling) with the first-order upper bound given in (2) and the componentwise upper bound given in (3). A linear regression model with two covariates and a sample size of 50 was used. Simulations were conducted conditional on the set of observed variables ($\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}$). The first observed covariate $\mathbf{x}_1$ took on equally spaced values in the interval (0, 100]. The second observed covariate was manipulated to produce a range of condition numbers for the X^TX matrix, but maintaining $\|\mathbf{x}_1\|_2 \approx \|\mathbf{x}_2\|_2$. The condition numbers examined were 1.18, 2.66, 12.37, and 72.29. The observed response variable $\mathbf{y}$ was generated from the model

$$y_i^R = 50 + 5x_{i1}^R + 10x_{i2}^R + \varepsilon_i.$$

The equation error ε_i was generated from a normal distribution with mean zero and variance σ^2. The variance was chosen to correspond to an R^2 of either .30 or .80. The statistic $R^2 = 1 - \|\mathbf{r}^x\|_2^2/\|\mathbf{y}\|_2^2$ is one minus the standardized length of the residual vector $\mathbf{r}^x$, and is a measure of the strength of the observed regression relationship. For each simulation, the random component was the measurement error (u_{ij}^R), which was uniformly distributed in the interval $(-a_j, a_j)$. In the first set of simulations given in Table 1(a), $a_1 = a_2 = 1$. For the second set of simulations given in Table 1(b), we assumed that there was measurement error only in the second covariate, $a_1 = 0$ and $a_2 = 1$. These errors are commensurate with the scale of the observed covariates, which are recorded in units of ones.

To calculate the first-order upper bound (2), we applied the methods of Davies and Hutton [5], Hodges and Moore [13], and Stewart [17] to estimate the unobserved $\|U\|_2$. In these papers, it is assumed that the measurement errors u_{ij}^R are independent, and that as the sample size n gets large, the matrix $n^{-1}U^TU$ converges in probability to the variance-covariance matrix of the measurement errors $\Lambda = \operatorname{diag}\{\lambda_j\}$. For the simulations in this paper, it is assumed that u_{ij}^R are uniformly distributed with support $[-a_j, a_j]$, and therefore that $\|U\|_2^2$ is estimated by max $(na_j^2/3)[j = 1, 2]$.

Tables 1(a) and 1(b) present the results of the simulations. These results compare the estimated upper bounds to the maximum relative error based on 10,000 replications. The two entries in each group represent the results for the two regression coefficients b_1, b_2. The upper bounds for the three methods are recorded for each regression coefficient, along with the maximum relative error based on 10,000 runs. The following observations are made:

(a) The diagnostic q (Corollary 1), which measures the size of the measurement error matrix U relative to the smallest eigenvalue of X^TX, is the primary indicator in

TABLE 1 (a)
$a_1 = 1, a_2 = 1.$

Condition number $\kappa(X'X)$	Maximum relative error	Stochastic upper bound	First-order upper bound	Componentwise upper bound
		$r^2 = .80$		
$\kappa(X'X) = 1.18$	.0292	.0768	.0721	.1301
$q = .031$	.0139	.0214	.0337	.0675
$\kappa(X'X) = 2.26$	.0312	.0809	.0708	.1362
$q = .040$	.0147	.0182	.0342	.0754
$\kappa(X'X) = 12.37$	.0595	.1992	.1545	.2988
$q = .083$	.0313	.0798	.0822	.1729
$\kappa(X'X) = 72.29$	.1181	.4698	.2626	.5567
$q = .117$	.0518	.1963	.1205	.2628
		$r^2 = .30$		
$\kappa(X'X) = 1.18$	.0624	.1637	.1493	.2416
$q = .031$	.0225	.0582	.0618	.1062
$\kappa(X'X) = 2.66$	.0385	.1056	.0955	.1731
$q = .040$	.0279	.0781	.0730	.1443
$\kappa(X'X) = 12.37$	.0775	.2193	.1950	.3931
$q = .083$	.0603	.1658	.1557	.3285
$\kappa(X'X) = 72.29$	.1795	.5078	.4059	.8420
$q = .117$	.1329	.3611	.2921	.6153

TABLE 1 (b)
$a_1 = 0, a_2 = 1.$

Condition number $\kappa(X'X)$	Maximum relative error	Stochastic upper bound	First-order upper bound	Componentwise upper bound
		$r^2 = .80$		
$\kappa(X'X) = 1.18$	.0263	.0580	.0721	.0715
$q = .022$	.0124	.0160	.0337	.0517
$\kappa(X'X) = 2.26$	.0233	.0577	.0708	.0765
$q = .028$	.0123	.0127	.0342	.0505
$\kappa(X'X) = 12.37$	.0533	.1555	.1545	.1823
$q = .059$	.0291	.0615	.0822	.1117
$\kappa(X'X) = 72.29$	.0799	.3312	.2626	.3205
$q = .083$	.0394	.1367	.1205	.1560
		$r^2 = .30$		
$\kappa(X'X) = 1.18$	.0285	.1238	.1493	.0828
$q = .022$	.0228	.0439	.0618	.0926
$\kappa(X'X) = 2.66$	.0170	.0755	.0955	.0710
$q = .028$	.0231	.0557	.0730	.0898
$\kappa(X'X) = 12.37$	.0581	.1712	.1950	.2083
$q = .059$	.0498	.1292	.1557	.1917
$\kappa(X'X) = 72.29$	.1155	.3582	.4059	.4175
$q = .083$	.0884	.2539	.2921	.3203

determining how badly the observed regression coefficients can be perturbed. It is not the size of the measurement errors or the condition number alone, but their joint relationship that determines the maximum relative error. This observation has been noted by Davies and Hutton [5], Beaton, Rubin, and Barone [1], and Stewart [16], [17]. A general rule is that if $q < .10$, the analyst can proceed, ignoring the measurement error.

(b) For $q < .10$ and $a_1 = a_2$ (Table 1(a)), the stochastic upper bound and the first-order upper bound are comparable. The componentwise upper bound is uniformly weaker over all entries of Table 1(a).

(c) For $q < .10$ and $a_1 \neq a_2$ (Table 1(b)), the stochastic upper bound is generally the tightest, with the first-order and componentwise upper bounds comparable but weaker.

(d) The first-order upper bound is at its strongest when $a_1 = a_2$, as is the case of the simulations presented in Table 1(a). This is a result of letting $\|U\|_2^2 = \max(na_j^2/3)$. However, when the magnitude of the measurement errors varies between covariates ($a_1 \neq a_2$), the bound may be considerably weakened, since only the maximum a_j is used in the calculation. This is illustrated in Table 1(b), where the first-order upper bound remained unchanged from Table 1(a), even though there was no longer measurement error in the first covariate. Note that both the stochastic and componentwise upper bounds are reduced in Table 1(b) as a result of setting $a_1 = 0$. In general, as the ratio $(\max a_j)/(\min a_j)$ increases away from one, the first-order upper bound gets weaker compared to the stochastic and componentwise upper bounds.

(e) The distance of the stochastic upper bound to the maximum relative error is relatively constant across values of q and R^2. The stochastic upper bound is approximately two to three times the size of the maximum relative error for most entries in Tables 1(a) and 1(b).

(f) The strength of the observed regression relationship affects the maximum relative error. A weaker regression implies that the regression coefficients take on a greater range of values, resulting in a larger maximum relative error. This is in line with first-order theory, which shows that the bound is of order $\kappa(X^TX)$ when the residual $\mathbf{r}^x$ is significant, and of order $\kappa(X)$ when $\mathbf{r}^x$ is small. The condition number for the matrix X is defined as $\kappa(X)$.

6. Concluding remarks. We have presented a diagnostic q and a stochastic upper bound that together describe the conditions under which the effect of measurement can be ignored for a given set of data. The resultant methodology encompasses a general measurement error framework, requires relatively simple computation, and can be applied to small data sets. We propose that an analyst examine the potential effects of measurement error. If the effects are minimal, then the presence of measurement error should cause no concern.

Results of the simulation study comparing this upper bound to a first-order and componentwise upper bound show that if the diagnostic q is less than .10, the stochastic upper bound is stable in the sense that it compares favorably to the other bounds over all conditions described in the simulation. In contrast, we found that while the first-order upper bound was at its best when $a_1 = a_2$, it produced a weaker bound relative to the stochastic upper bound when the precision of the measurement error varied between covariates. We also found that the componentwise upper bound was greater than the stochastic upper bound over almost all entries in Tables 1(a) and 1(b).

The method of least squares generally produces poor estimates of the regression coefficients when the covariates are nearly collinear. Our simulations have demonstrated that the stochastic upper bound for the change in the least squares estimate (as well as the other bounds studied) is weak when the condition number of the observed covariate matrix is large. The situations where the stochastic upper bound is weak correspond to situations where least squares estimation is unsatisfactory. Therefore, this methodology should not be applied in highly collinear situations. If the data contains nearly collinear covariates, other methods of estimating the regression coefficients should be considered, such as principal components, ridge regression, or the removal of covariates [4]. The computation of bounds for least squares estimates in these situations is moot.

In conclusion, we believe that the applicability of the presented methodology to most data sets will encourage data analysts to address the problem of measurement error, which is often ignored, in the linear regression model.

Appendix.

Proof of Theorem 1. The matrix norm used in this proof is the spectral norm (§ 1, equation (1a)). We begin with the equality

$$\text{(A1)}\qquad \begin{aligned} |\mathbf{g}_j^T\mathbf{y} - \mathbf{h}_j^T\mathbf{y}|^2 = [\{&\|P_z(\mathbf{g}_j - \mathbf{h}_j)\|^2 + \|(I - P_z)(\mathbf{g}_j - \mathbf{h}_j)\|^2\} \\ &\times \|\mathbf{y}\|^2\{\cos^2\theta(\mathbf{y}, \mathbf{g}_j - \mathbf{h}_j)\}], \end{aligned}$$

where

$$\begin{aligned} P_x &= X(X^TX)^{-1}X^T, & P_z &= Z(Z^TZ)^{-1}Z^T, \\ \mathbf{g}_j &= X(X^TX)^{-1}\mathbf{e}_j & \mathbf{h}_j &= Z(Z^TZ)^{-1}\mathbf{e}_j, \\ b_j^x &= \mathbf{g}_j^T\mathbf{y}, & b_j^z &= \mathbf{h}_j^T\mathbf{y}, \end{aligned}$$

and $\mathbf{e}_j$ is the jth unit basis vector. The identity in (A1) follows from the Pythagorean Theorem and the Law of Cosines. In addition, it is noted that $P_z\mathbf{h}_j = \mathbf{h}_j$ and $(I - P_z)\mathbf{h}_j = \mathbf{0}$.

To develop an upper bound for the difference between the observed and true jth estimated regression coefficients, a bound must be found for the first two terms inside the bracket and the cosine function on the right-hand side of (A1).

(i) A bound on the second term inside the bracket is

$$\text{(A2)} \qquad \|(I-P_z)(\mathbf{g}_\mathbf{j}-\mathbf{h}_\mathbf{j})\|^2 = \|(I-P_z)(Z+U)X^+\mathbf{g}_\mathbf{j}\|^2 \leqq \|\mathbf{g}_\mathbf{j}\|^2\|X^+\|^2\|U\|^2,$$

where the equality follows from $(I-P_Z)\mathbf{h}_\mathbf{j} = \mathbf{0}$ and $P_x\mathbf{g}_\mathbf{j} = \mathbf{g}_\mathbf{j}$.

(ii) To calculate an upper bound for $\|P_Z(\mathbf{g}_\mathbf{j}-\mathbf{h}_\mathbf{j})\|^2$, we use the assumption stated in Theorem 1,

$$\text{(A3)} \qquad \|P_z\mathbf{g}_j\|^2 \leqq \|\mathbf{h}_j\|^2 \leqq \|\mathbf{g}_j\|^2[1 + (\|\mathbf{h}_j - \mathbf{g}_j\|^2/\|\mathbf{g}_j\|^2)].$$

This assumption restricts the size of the unobservable vector $\mathbf{h}_j$ to lie in the neighborhood of the observed $\|\mathbf{g}_j\|$.

Letting $d = \|X^+\|\,\|U\|$, it follows from Fig. 1 and (A2) that

$$\sin\theta_1 = \|(I-P_z)\mathbf{g}_\mathbf{j}\|/\|\mathbf{g}_\mathbf{j}\| \leqq d.$$

In addition, since θ_1 and θ_3 are complementary angles,

$$\text{(A4)} \qquad \theta_3 \geqq \pi/2 - \sin^{-1}(d).$$

Now, based on assumption (A3),

$$\cos(\theta_3+\theta_4) = \frac{\|\mathbf{g}_j\|^2 + \|\mathbf{g}_j - \mathbf{h}_j\|^2 - \|\mathbf{h}_j\|^2}{2\|\mathbf{g}_j\|\,\|\mathbf{g}_j - \mathbf{h}_j\|} \geqq 0$$

or $\theta_3 + \theta_4 \leqq \pi/2$. Therefore, by (A4),

$$\theta_4 \leqq \sin^{-1}(d),$$

and since θ_2 and θ_4 are complementary angles,

$$\theta_2 > \pi/2 - \sin^{-1}(d) \quad \text{and} \quad \cos(\theta_2) \leqq d.$$

If $\cos\theta_2$ is rewritten as $\|P_Z(\mathbf{g}_\mathbf{j}-\mathbf{h}_\mathbf{j})\|/\|\mathbf{g}_\mathbf{j}-\mathbf{h}_\mathbf{j}\|$ (see Fig. 1), then a bound on $\|P_Z(\mathbf{g}_\mathbf{j}-\mathbf{h}_\mathbf{j})\|$ will be accomplished by bounding $\|\mathbf{g}_\mathbf{j}-\mathbf{h}_\mathbf{j}\|$. Using a standard decom-

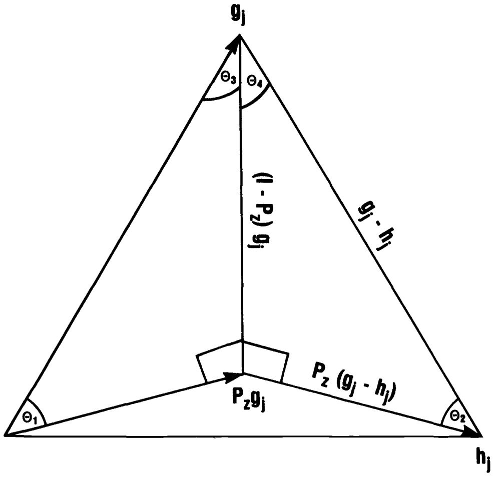


FIG. 1

position in perturbation theory, $\mathbf{g}_j^T - \mathbf{h}_j^T = \mathbf{g}_j^T UZ^+ - \mathbf{g}_j^T P_x(I - P_z)$, and assuming $\|X^+\| \|U\| < 1$ (see, for example, [19]):

$$\|\mathbf{g}_j - \mathbf{h}_j\| \leqq \|\mathbf{g}_j\| d\{(2-d)/(1-d)\}.$$

Since $\cos(\theta_2) \leqq d$, it follows that

$$\text{(A5)} \qquad \|P_z(\mathbf{g}_j - \mathbf{h}_j)\| \leqq \|\mathbf{g}_j\| d^2\{(2-d)/(1-d)\}.$$

(iii) To obtain a bound on the cosine function in (A1), let $\theta_5 = \theta_3 + \theta_4$, θ_6 be the angle formed by $\mathbf{g}_j$ and $\mathbf{y}$, and θ_7 be the angle formed by $\mathbf{g}_j - \mathbf{h}_j$ and $\mathbf{y}$.

To bound $|\cos\theta_7|$, we note that $\theta_5 \leqq \theta_6 + \theta_7$ and, since $\theta_3 \geqq \pi/2 - \sin^{-1}(d)$, it follows that $\pi/2 \geqq \theta_5 \geqq \cos^{-1}(d)$. So

$$\theta_7 \geqq \cos^{-1}(d) - \theta_6 \quad \text{and} \quad \cos\theta_7 \leqq \cos\{\cos^{-1}(d) - \theta_6\}.$$

Therefore,

$$\text{(A6)} \qquad |\cos\theta(\mathbf{g}_j - \mathbf{h}_j)\mathbf{y}| \leqq \cos\{\cos^{-1}(d) - \theta_6\}.$$

(If $\{\cos^{-1}(d) - \theta_6\} < 0$, then we set the upper bound of (A6) to 1.) We use the notation Γ to denote the upper bound in (A6), and γ to denote the angle θ_6.

Therefore, from (A1), (A2), (A5), and (A6),

$$\frac{|b_j^x - b_j^z|}{|b_j^x|} \leqq \frac{\Gamma d[\{d(2-d)\}^2 + (1-d)^2]^{1/2}}{|\cos\gamma|(1-d)}.$$

Proof of Lemma 1. Let $f_W(w) = \frac{1}{2a}$ and $-a \leqq W \leqq a$. Let $S = V^2$, where the density of V satisfies the conditions in Lemma 1. If F is the distribution function of S, we want to show that $\inf_f F_s(s)$ occurs when f is the uniform density. Note that

$$F_s(s) = 2\int_0^{\sqrt{s}} f_V(v)\,dv$$

by symmetry about zero. Since $(1/\sqrt{s})F_s(s)$ is a decreasing function of $\sqrt{s}$,

$$2\int_0^{\sqrt{s}} f_V(v)\,dv > \frac{\sqrt{s}}{a} \qquad (0 < \sqrt{s} < a).$$

But $\sqrt{s}/a = 2\int_0^{\sqrt{s}} f_W(w)\,dw$, i.e., the minimizing distribution is uniform.

Proof of Theorem 2. Proof of Theorem 2 begins with

$$\text{(A7)} \qquad \phi(D^{-1}B^+) \leqq \|D^{-1}B^+\|_2 \leqq \sqrt{p}\,\phi(D^{-1}B^+)$$

(see, for example, [10, pp. 15–16]). Given $D = \operatorname{diag}\{d_j\}$, if $d_j \leqq 1$, $j = 1, 2, \ldots, p$, $\phi(B^+) \leqq \phi(D^{-1}B^+)$, and therefore, $\phi(B^+) \leqq \|D^{-1}B^+\|_2$. Now

$$\phi(\tilde{D}^{-1}B^+) = \tilde{d}_j^{-1}\|e_j^T B^+\|_2 \qquad j = 1, 2, \ldots, p,$$

and since $\tilde{D}$ is chosen so that the rows of B^+ are of equal length with $\tilde{d}_j = 1$ for some j,

$$\phi(\tilde{D}^{-1}B) = \phi(B^+).$$

This proves (i), and (ii) follows immediately from (A7).

REFERENCES

[1] A. E. Beaton, D. B. Rubin, and J. L. Barone, *The acceptability of regression solutions: Another look at computational accuracy*, J. Amer. Statist. Assoc., 71 (1976), pp. 158–168.

[2] R. J. CARROLL, P. P. GALLO, AND L. J. GLESER, *Comparison of least squares and errors-in-variables regression, with special reference to randomized analysis of covariance*, J. Amer. Statist. Assoc., 80 (1985), pp. 929–932.

[3] J. M. CHAMBERS, W. S. CLEVELAND, B. KLEINER, AND P. A. TUKEY, *Graphical Methods for Data Analysis*, Duxbury Press, Boston, 1983.

[4] S. CHATTERJEE AND B. PRICE, *Regression Analysis by Example*, John Wiley and Sons, New York, 1991.

[5] R. B. DAVIES AND B. HUTTON, *The effects of errors in the independent variables in linear regression*, Biometrika, 62 (1975), pp. 383–391.

[6] R. FLETCHER, *Expected conditioning*, IMA J. Numer. Anal., 5 (1985), pp. 247–273.

[7] W. A. FULLER, *Measurement Error Models*, J. Wiley and Sons, New York, 1987.

[8] P. P. GALLO, *Consistency of regression estimates when some variables are subject to error*, Comm. Statist. Theory Methods, 9 (1982), pp. 973–983.

[9] L. J. GLESER, R. J. CARROLL, AND P. P. GALLO, *The limiting distribution of least squares in an error-in-variables regression model*, Ann. Statist., 15 (1987), pp. 220–233.

[10] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[11] N. J. HIGHAM, *Computing error bounds for regression problems, statistical analysis of measurement error models and applications*, Contemp. Math. 112, P. J. Brown and W. A. Fuller, eds., Amer. Math. Soc., Providence, RI, 1990, pp. 195–208.

[12] N. J. HIGHAM AND G. W. STEWART, *Numerical Linear Algebra in Statistical Computing, The State of the Art in Numerical Analysis*, A. Iserles and M. J. D. Powell, eds., Oxford University Press, London, pp. 41–57.

[13] S. D. HODGES AND P. G. MOORE, *Data uncertainties and least squares regression*, Appl. Statist., 21 (1972), pp. 185–195.

[14] S. M. ROSS, *Stochastic Processes*, John Wiley and Sons, New York, 1983.

[15] G. W. STEWART, *Sensitivity coefficients for the effects of errors in the independent variables in a linear regression*, Tech. Rep., TR-571, Dept. of Computer Science and Institute for Physical Science and Technology, Univ. of Maryland, Baltimore, 1977.

[16] ———, *Collinearity and least squares regression* (*with discussion*), Statist. Sci., 2 (1987), pp. 68–100.

[17] ———, *Perturbation theory and least squares with errors in the variables, statistical analysis of measurement error models and applications*, Contemp. Math., 112, P. J. Brown and W. A. Fuller, eds., Amer. Math. Soc., Providence, RI, 1990, pp. 171–181.

[18] A. VAN DER SLUIS, *Condition numbers and equilibration matrices*, Numer. Math., 14 (1973), pp. 14–23.

[19] P.-A. WEDIN, *Perturbation theory for pseudo-inverses*, BIT, 13 (1973), pp. 217–232.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 688–704, July 1993

007

MIXED, COMPONENTWISE, AND STRUCTURED CONDITION NUMBERS*

I. GOHBERG† AND I. KOLTRACHT‡

Abstract. Mixed, componentwise, and structured condition numbers for continuous maps in finite-dimensional spaces are introduced. For differentiable maps, convenient formulas for the estimation of such condition numbers are given and illustrated on the example of Toeplitz matrices. Applications to Vandermonde matrices are also presented.

Key words. condition number, componentwise error, structure, Toeplitz matrix, Vandermonde matrix

AMS subject classifications. 15A09, 65F35, 65G05

1. Introduction. In this paper, we study condition numbers of maps in finite-dimensional spaces $F : \mathbb{R}^p \to \mathbb{R}^q$. The condition number of F at a point $a \in D_F$ characterizes the instantaneous rate of change in Fa with respect to perturbations in $\mathbf{a}$. It provides an approximate error bound for computing Fa in the worst case over all perturbations. Definition and properties of a condition number can be found in papers by Lyubich [14] and Rice [15]. The usual way of measuring perturbations in $\mathbf{a}$ and Fa is by the relative error in norm. Thus the usual condition number relates $\|Fx - Fa\| / \|Fa\|$ to $\|x - a\| / \|a\|$. If F is differentiable at $\mathbf{a}$ then the usual condition number of F at $\mathbf{a}$ is

$$k(F, a) = \frac{\|F'(a)\| \|a\|}{\|Fa\|},$$

where $F'(a) : \mathbb{R}^p \to \mathbb{R}^q$ is the derivative of F at $\mathbf{a}$.

Skeel [17] defines a new condition number for the solution of linear systems of equations, $Ay = b$, which relates $\|Fx - Fa\| / \|Fa\|$ to perturbations in $\mathbf{a}$ satisfying the componentwise inequality

$$(1.1) \qquad |x_i - a_i| < \varepsilon |a_i|, \qquad i = 1, \dots, p.$$

Skeel's condition number characterizes perturbations in y with respect to perturbations in A and b caused by their representation in a computer arithmetic, which satisfies (1.1). It gives more accurate error bounds than the usual condition number. It is also useful for the analysis of sparse matrices (see also Arioli, Demmel, and Duff [1], Demmel and Kahan [6], and Higham [11]).

More accurate error bounds for perturbations in Fa can also be obtained if it is known that perturbations of $\mathbf{a}$ are restricted to a proper subset S of D_F, which is parameterized by a map from a space $\mathbb{R}^n$ to $\mathbb{R}^p$ with $n < p$. Such bounds for the solution of $Ay = b$, where A is a Vandermonde matrix, are given by Higham [10]. Analysis of the inversion of Cauchy matrices is given in Gohberg and Koltracht [8]. (See also Babuška [2], Van Dooren [20], and Higham and Higham [12].)

In this paper we propose condition numbers for a general continuous map F at a point $\mathbf{a}$, which relate normed or componentwise errors in Fa to componentwise errors

* Received by the editors December 7, 1990; accepted for publication (in revised form) December 17, 1991. This work was supported in part by National Science Foundation grant DMS-9007030.

† Department of Mathematics, Tel Aviv University, Ramat Aviv, 69978, Tel Aviv, Israel (gohberg@taurus).

‡ Department of Mathematics, University of Connecticut, Storrs, Connecticut 06269-0009 (koltrach@uconnvm).

in **a**, as well as to perturbations in **a** restricted to a "low-dimensional" structured subset S. Such a structured subset consisting of Vandermonde matrices is considered in § 3.

In § 2.1, we consider the definition and properties of the usual condition number. In § 2.2, we introduce *mixed* and *componentwise* condition numbers, which relate normed and componentwise errors in Fa, respectively, to componentwise errors in **a**.

In §§ 2.3 and 2.4 we give expressions of mixed and componentwise condition numbers for two concrete problems, matrix inversion, and solution of a linear system of equations, respectively. In § 2.5, we introduce the notion of a *structured* condition number and discuss its relevance in connection with the numerical stability of algorithms for structured computational problems, with examples of Cauchy and Toeplitz matrices.

In § 3.1 we give an a priori bound for the structured mixed condition number of the map of matrix inversion on the subset of Vandermonde matrices $A = (a_j^i)_{i,j=0}^n$. We show that if $F : (a_0, \dots, a_n) \rightarrow y$, where y is any row of A^{-1}, then

$$m(F, a) \leqq n^2 \max (n\Delta, n + \Delta)$$

where

$$\Delta = \max_{\substack{i,j=0,\dots,n \\ i \neq j}} \frac{|a_i|}{|a_i - a_j|}.$$

This a priori bound indicates that the structured condition number of a row of the inverse of a Vandermonde matrix can be much smaller than the usual condition number $\|A\| \, \|A^{-1}\|$, as the latter one can grow exponentially with n (see Gautschi and Inglese [7]). This suggests that we should look for algorithms that take full account of the structure of a Vandermonde matrix. Two such algorithms, those of Björck and Pereyra and of Traub, are shown in certain cases to yield errors that are not a multiple of the structured condition number when they are used to compute the inverse of a Vandermonde matrix. We present a new inversion algorithm that reflects the structured condition number in numerical experiments, and we conjecture that it is stable in this sense in general. In § 3.2, we show how this new algorithm can be used for an efficient solution of linear systems of equations with Vandermonde coefficient matrices.

2. Condition numbers.

In this section, we consider condition numbers of a continuous map from one finite-dimensional space to another.

2.1. The usual condition number.

Let $F : \mathbb{R}^p \rightarrow \mathbb{R}^q$ be a continuous map in normed linear spaces defined on an open set $D_F \subset \mathbb{R}^p$. For a given $a \in D_F$, $a \neq 0$, such that $Fa \neq 0$ and $\varepsilon > 0$ small enough such that $B(a, \varepsilon) = \{x \in \mathbb{R}^p : \|x - a\| \leqq \varepsilon\} \subset D_F$, the ε condition number $k(F, a, \varepsilon)$ is defined by

$$k(F, a, \varepsilon) = \sup_{\substack{x \in B(a,\varepsilon) \\ x \neq a}} \left\{ \frac{\|Fx - Fa\|}{\|Fa\|} \Big/ \frac{\|x - a\|}{\|a\|} \right\}. \tag{2.1}$$

Although $\mathbb{R}^p$ and $\mathbb{R}^q$ may be equipped with different norms, we will use the same notation for both of them as long as it does not cause any confusion. The number $k(F, a, \varepsilon)$ gives the worst possible magnification of errors in **a** (of size at most ε) when **a** is transformed into Fa. Since $k(F, a, \varepsilon)$ is nonincreasing when ε goes to zero, there exists

$$\lim_{\varepsilon \rightarrow 0} k(F, a, \varepsilon) = k(F, a).$$

By definition, the number $k(F, a)$ is called the condition number of the map F at the point **a**. This definition appears in Lyubich [14] and Rice [15]. If F is differentiable at **a**, then it is easy to see that

$$(2.2)\qquad k(F,a)=\frac{\|F'(a)\|\,\|a\|}{\|Fa\|},$$

where $\|F'(a)\|$ is the norm of the derivative of F at **a** as a linear map $F'(a):\mathbb{R}^p\rightarrow\mathbb{R}^q$.

One purpose of the condition number is to provide approximate error bounds. For x sufficiently close to **a** we have

$$(2.3)\qquad \frac{\|Fx-Fa\|}{\|Fa\|}\leqq k(F,a)\frac{\|x-a\|}{\|a\|}+o(\|x-a\|).$$

If, say, Fa is to be computed in finite arithmetic with unit round-off error u, then the representation $fl(a)$ of **a** in this arithmetic satisfies $\|fl(a)-a\|_\infty\leqq u\|a\|_\infty$. Thus, even without the effect of round-off errors in the performed calculation, the error in the computed value Fa can be of size $k(F,a)u$. The problem of computing the value of the map F at the point **a** is said to be well or ill conditioned depending on how large $k(F,a)u$ is relative to the allowed error in the computed value Fa. (In the context of matrix computations, see Wilkinson [21], Golub and Van Loan [9], and Bunch [5] for more details.)

There are situations in which the usual condition number, $k(F,a)$, may give unnecessarily large error bounds. Next, we consider cases in which more adequate condition numbers can be defined.

2.2. Mixed and componentwise condition numbers. Here we assume that perturbations x of the data point **a** satisfy the componentwise inequality (1.1). Note that the zero components of **a** are not perturbed. This is the case, for example, when $x_i=fl(a_i)$ is the representation of the component a_i in a computer arithmetic. Thus $|x_i-a_i|\leqq u|a_i|$, $i=1,\dots,p$, where u is the unit round-off error. It is clear that the inequality (1.1) implies that

$$(2.4)\qquad \|x-a\|_\infty\leqq\varepsilon\|a\|_\infty.$$

This is, of course, not true in the opposite direction. In fact, if (2.4) holds, then (1.1) can be guaranteed for the largest component of **a** only, namely, $|x_{i_0}-a_{i_0}|\leqq\varepsilon|a_{i_0}|$, where $|a_{i_0}|=\|a\|_\infty$. Relative errors in other components of **a** can be arbitrary. Therefore, the problem of computing F at **a** could be ill conditioned with respect to perturbations in **a** satisfying (2.4) while being well conditioned with respect to perturbations satisfying (1.1). Here we propose a definition of a *mixed* condition number, which relates errors in Fa in norm to componentwise perturbations in **a**. Let

$$B^0(a,\varepsilon)=\{x:|x_i-a_i|<\varepsilon|a_i|,\ i=1,\dots,p\}.$$

DEFINITION 2.1. Let $F:\mathbb{R}^p\rightarrow\mathbb{R}^q$ be a continuous map defined on an open set $D_F\subset\mathbb{R}^p$. For a given $a\in D_F$, $a\neq 0$, such that $Fa\neq 0$ and $\varepsilon>0$ small enough such that $B^0(a,\varepsilon)\subset D_F$, let

$$m(F,a,\varepsilon)=\sup_{\substack{x\in B^0(a,\varepsilon)\\ x\neq a}}\left\{\frac{\|Fx-Fa\|_\infty}{\|Fa\|_\infty}\Big/\delta(x,a)\right\}$$

where

$$(2.5)\qquad \delta(x,a)=\max_{\substack{i=1,\dots,p\\ a_i\neq 0}}\left\{\frac{|x_i-a_i|}{|a_i|}\right\}.$$

The mixed condition number of the map F at the point $\mathbf{a}$ is

$$m(F, a) = \lim_{\varepsilon \to 0} m(F, a, \varepsilon).$$

Note that $B^0(a, \varepsilon) \subset B(a, \varepsilon\|a\|_\infty)$ and that for $x \in B^0(a, \varepsilon)$

$$\frac{\|x-a\|_\infty}{\|a\|_\infty} \leqq \max_{\substack{i=1,\dots,p \\ a_i \neq 0}} \left\{ \frac{|x_i - a_i|}{|a_i|} \right\} = \delta(x, a).$$

Therefore, $m(F, a, \varepsilon) \leqq k(F, a, \varepsilon\|a\|_\infty)$, and hence,

$$m(F, a) \leqq k(F, a),$$

where $k(F, a)$ is taken with respect to the infinity norm. For the map $F : (A, b) \to y$, where y is the solution of a linear system of equations $Ay = b$, such a mixed condition number is given by Skeel [17, Thm. 2.3], namely,

$$m(F, (A, b)) = \frac{\| |A^{-1}| |A| |y| + |A^{-1}| |b| \|_\infty}{\|y\|_\infty}, \tag{2.6}$$

where $|A| = \{|a_{ij}|\}_{i,j=1}^n$.

If F is differentiable at $\mathbf{a}$, then we can show (using, for example, the same argument as in Gohberg and Koltracht [8]) that

$$m(F, a) = \frac{\|F'(a)D_a\|_\infty}{\|Fa\|_\infty}. \tag{2.7}$$

In this case, if $x \in B^0(a, \varepsilon)$ with ε sufficiently small, then

$$\frac{\|Fx - Fa\|_\infty}{\|Fa\|_\infty} \leqq \frac{\|F'(a)D_a\|_\infty}{\|Fa\|_\infty} \varepsilon + o(\varepsilon).$$

Following Gohberg and Koltracht [8] we also give the definition of a *componentwise* condition number, which relates componentwise errors in Fa to componentwise perturbations in $\mathbf{a}$. This condition number gives a more detailed characterization of errors in Fa and is useful when Fa is the data point for another map.

DEFINITION 2.2. Under the conditions of Definition 2.1 let $Fa = (f_1(a), \dots, f_q(a))$ be such that $f_j(a) \neq 0, j = 1, \dots, q$. Then the componentwise condition number of the map F at the point $\mathbf{a}$ is

$$c(F, a) = \lim_{\varepsilon \to 0} \sup_{\substack{x \in B^0(a,\varepsilon) \\ x \neq a}} \{\delta(Fx, Fa)/\delta(x, a)\},$$

where

$$\delta(Fx, Fa) = \max_{j=1,\dots,q} \frac{|f_j(x) - f_j(a)|}{|f_j(a)|}.$$

It is shown in Gohberg and Koltracht [8] that if F is differentiable at a, then

$$c(F, a) = \|D_{Fa}^{-1} F'(a) D_a\|_\infty \tag{2.8}$$

where $D_{Fa} = \operatorname{diag}\{f_1(a), \dots, f_q(a)\}$. If $x \in B^0(a, \varepsilon)$ with ε sufficiently small, then

$$\delta(Fx, Fa) \leqq \|D_{Fa}^{-1} F'(a) D_a\|_\infty \varepsilon + o(\varepsilon).$$

It is straightforward to check that the condition numbers k, m, and c satisfy the following semiadditivity relations.

Let $F = F_2 \circ F_1$ where $F_1 : \mathbb{R}^n \rightarrow \mathbb{R}^p$ and $F_2 : \mathbb{R}^p \rightarrow \mathbb{R}^q$ are continuous maps. Let $a \in D_{F_1}$, $F_1 a \in D_{F_2}$, $a \neq 0$, $F_1 a \neq 0$ and $Fa \neq 0$. Then

(i) $k(F, a) \leqq k(F_1, a) k(F_2, F_1 a)$.

(ii) $m(F, a) \leqq m(F_1, a) k(F_2, F_1 a)$.

(iii) $m(F, a) \leqq c(F_1, a) m(F_2, F_1 a)$ provided that $F_1 a$ has no zero components.

(iv) $c(F, a) \leqq c(F_1, a) c(F_2, F_1 a)$ provided that $F_1 a$ and $F_2 \circ F_1 a$ have no zero components.

These relations indicate the limitations of the mixed condition number $m(F, a)$, namely, it is applicable to the perturbation of original data, or to a value of another map with no zero components, only. Next we consider two particular maps.

2.3. Matrix inversion. Let F be the map of matrix inversion defined on the open set of all nonsingular $n \times n$ matrices. Then $F'(A)(\cdot) = -A^{-1}(\cdot)A^{-1}$ and $k(F, A) = \|A\| \|A^{-1}\|$ (see, for example, Belitskii and Lyubich [3]). To find $m(F, A)$, consider F as a map from $\mathbb{R}^{n^2}$ to $\mathbb{R}^{n^2}$, where the data vector consists of all elements of A arranged, say, row by row and so is $FA = A^{-1}$. Let $A^{-1} = \{\sigma_{ij}\}_{i,j=1}^{n}$ and observe that

$$\frac{\partial \sigma_{ij}}{\partial a_{pq}} = -\sigma_{ip}\sigma_{qj}, \qquad i, j, p, q = 1, \ldots, n.$$

In particular, the infinity norm of A^{-1} as a vector is

$$\max_{i,j=1,\ldots,n} |\sigma_{ij}| \triangleq \|A^{-1}\|_\nu.$$

The $((n-1)i + j)$th row of $F'(A)D_A$ (regarded as a matrix in $\mathbb{R}^{n^2 \times n^2}$) is of the form $[\sigma_{i1}a_{11}\sigma_{1j}, \sigma_{i1}a_{12}\sigma_{2j}, \ldots, \sigma_{in}a_{nn}\sigma_{nj}]$, and its 1-norm is therefore $|r_i|\,|A|\,|c_j|$, where r_i is the ith row of A^{-1} and c_j is the jth column of A^{-1}. Thus

$$m(F, A) = \frac{\|F'(A)D_A\|_\infty}{\|FA\|_\infty} = \frac{\|\,|A^{-1}|\,|A|\,|A^{-1}|\,\|_\nu}{\|A^{-1}\|_\nu}.$$

It is easy to see that if A^{-1} has no zero entries, then

$$c(F, A) = \left\| \left(\frac{\alpha_{ij}}{\sigma_{ij}} \right)_{i,j=1}^{n} \right\|_\nu,$$

where $|A^{-1}|\,|A|\,|A^{-1}| = (\alpha_{ij})_{i,j=1}^{n}$ (see also Rohn [16]). Note that if C_k maps A into c_k, the kth column of A^{-1}, then

$$m(C_k, A) = \frac{\|\,|A^{-1}|\,|A|\,|c_k|\,\|_\infty}{\|c_k\|_\infty}, \tag{2.9}$$

and if R_k maps A into r_k, the kth row of A^{-1}, then

$$m(R_k, A) = \frac{\|\,|r_k|\,|A|\,|A^{-1}|\,\|_\infty}{\|r_k\|_\infty}. \tag{2.10}$$

If the column c_k or the row r_k have no zero components, then, respectively,

$$c(C_k, A) = \|D_{c_k}^{-1}\,|A^{-1}|\,|A|\,|c_k|\,\|_\infty$$

and

$$c(R_k, A) = \|\,|r_k|\,|A|\,|A^{-1}|D_{r_k}^{-1}\|_\infty.$$

It is clear that $m(C_k, A)$ does not depend on the row scaling of A, and $m(R_k, A)$ does not depend on the column scaling of A, (see also Skeel [17]). Indeed, if $T =$

diag $\{t_1, \ldots, t_n\}$ and $S = \text{diag}\,\{s_1, \ldots, s_n\}$, then

$$m(C_k, TA) = m(C_k, A), \qquad m(R_k, AS) = m(R_k, A).$$

It is interesting to note that the componentwise condition number does not depend on either column or row scaling. That is,

$$c(C_k, TAS) = c(C_k, A)$$

and

$$c(R_k, TAS) = c(R_k, A).$$

2.4. Solution of $Ax = b$. Let $F : (A, b) \to y$, where y is the solution of the linear system of equations $Ay = b$. Then it is well known that

$$\sup_{b \in \mathbb{R}^n} k(F, (A, b)) = 2\|A\|\,\|A^{-1}\|.$$

To find the mixed condition number observe that $Ay = b$ is equivalent to

$$\begin{bmatrix} A & -b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

It immediately follows from (2.9) with $k = n + 1$ that

$$m(F, (A, b)) = \frac{\|\,|A^{-1}|\,|A|\,|y| + |A^{-1}|\,|b|\,\|_\infty}{\|y\|_\infty},$$

which is Skeel's condition number (2.6). If y has no zero components, then

$$c(F, (A, b)) = \|D_y^{-1}(|A^{-1}|\,|A|\,|y| + |A^{-1}|\,|b|)\|_\infty.$$

Once again, it is clear that $m(F, (A, b))$ does not depend on row scaling of A, while $c(F, (A, b))$ does not depend on either row or column scaling of A. (This does not mean, of course, that scaling cannot improve performance of a particular algorithm for solving $Ay = b$.)

2.5. Structured condition numbers. Suppose that a point b belongs to a subset $S \subset D_G$, where $G : \mathbb{R}^p \to \mathbb{R}^q$ and S is parameterized by a map H from a space the dimension of which is less than p. Furthermore, suppose that perturbations in b are restricted to the subset S. Then the errors in Gb can be smaller (sometimes much smaller) than predicted by (2.2). To obtain more adequate error bounds we give the following definition of a *structured* condition number.

DEFINITION 2.3. Let G be a continuous map $G : \mathbb{R}^p \to \mathbb{R}^q$. Let $H : \mathbb{R}^n \to \mathbb{R}^p$, $n < p$, be a continuous one-to-one map such that $HD_H = S \subset D_G$. Then the subset S is called a structured subset of D_G with the structure imposed by the map H. The corresponding structured condition number of G at $b \in S$ is by definition the condition number (k, m, or c) of $F = G \circ H$ at $a = H^{-1}b$.

The restriction that H is one-to-one can be omitted, in which case the structured condition number is not unique. In this case, it is important to find a preimage of b at which F has the smallest condition number.

Apart from providing error bounds for the initial perturbation of data, structured condition numbers can be useful for understanding numerical stability of algorithms for structured problems. Following the general definition in Stoer and Bulirsch [18, § 1.3], an algorithm for computing Fa is said to be stable if the contribution of intermediate round-off errors to the error in computed Fa is of the same order of magnitude as that of the original perturbation of data. This notion of stability is often referred to as *forward stability* and is not to be confused with the stronger property of *backward stability*, which requires that the computed solution is the exact solution of a slightly perturbed problem.

Let us consider now the example of matrix inversion. Let G be the map of matrix inversion defined on the set of all $n \times n$ nonsingular matrices. Let S be a structured subset of D_G (e.g., Toeplitz, Vandermonde, or Cauchy matrices), with the structure imposed by $H : \mathbb{R}^m \to \mathbb{R}^{n^2}$. Suppose that we have two algorithms for inverting $B \in S$. One is a general purpose matrix inversion algorithm defined on D_G, and the second is a structured algorithm for computing values of $F = G \circ H$ on $H^{-1}(S)$. Let the general condition number of G at B be much larger than the structured condition number (of the same type) of F at $a = H^{-1}(B)$. Then the general purpose algorithm, even if it is stable on D_G, may become unstable on the structured subset S. Indeed, a column of the computed inverse would be exact for some nonsingular matrix close to B, but may not be an exact column of the inverse of any matrix from S. In this case, the structured algorithm should be preferred if it is shown to be stable on $D_F = H^{-1}(S)$. Using forward error analysis, the stability of Björck–Pereyra algorithms on a subclass of Vandermonde matrices was established by Higham [10] and for an algorithm for the inversion of Cauchy matrices by Gohberg and Koltracht [8] (see also Example 1 below).

On the other hand, it may happen that there is not much difference between general and structured condition numbers (the structured condition number may be, in principle, even larger). In this case, a stable general purpose algorithm remains stable on the structured class and should be preferred unless it is shown that the structured algorithm is more trustworthy (for definition, see Stoer and Bulirsch [18, § 1.3]). We will demonstrate elsewhere that the Cholesky algorithm is stable on positive definite Toeplitz matrices and is more trustworthy than the Levinson algorithm. The relevance of structured perturbations to the numerical stability of algorithms is also discussed in Babuška [2], Van Dooren [20], and Higham and Higham [12].

Structured condition numbers can be found in different ways. If the values of Fa are given explicitly in terms of entries of $\mathbf{a}$, then we can compute the derivative of F and use (2.7) or (2.8). This can be done for Cauchy matrices (see Gohberg and Koltracht [8] and Example 1 below), and for Vandermonde matrices (see § 3, and for a different method, Higham [10]). If the values of Fa are not available in terms of $\mathbf{a}$, but the derivatives of G and H are known, then we can use the chain rule

$$F'(a) = G'(b)H'(a).$$

In Example 2 we show how to find $F'(a)$ for Toeplitz matrices when G is the map of matrix inversion or the map of a simple eigenvalue.

Example 1. Let G be the map of matrix inversion $G : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$, and let D_G be the open set of all $n \times n$ nonsingular matrices. Let S be a subset of D_G consisting of all Cauchy matrices

$$S = \left\{ \left(\frac{1}{t_i - s_j} \right)_{i,j=1}^{n} \right\},$$

where $(t_1, \ldots, t_n, s_1, \ldots, s_n)$ is an arbitrary $2n$-tuple of pairwise different numbers. Thus

$$H(t_1, \ldots, t_n, s_1, \ldots, s_n) = \left(\frac{1}{t_i - s_j} \right)_{i,j=1}^{n}, \qquad F = G \circ H : \mathbb{R}^{2n} \to \mathbb{R}^{n^2}$$

and

$$F(t_1, \ldots, t_n, s_1, \ldots, s_n) = \left[\left(\frac{1}{t_i - s_j}\right)_{i,j=1}^{n}\right]^{-1}.$$

It is shown in Gohberg and Koltracht [8] that

$$c(F, a) \leqq \frac{\|a\|_\infty}{\Delta}(8n - 6),$$

where

$$\Delta = \min_{\substack{i,j=1,\ldots,2n \\ i \neq j}} |a_i - a_j| \quad \text{with } a = (t_1, \ldots, t_n, s_1, \ldots, s_n) = (a_1, \ldots, a_{2n}).$$

In particular, for a Hilbert matrix

$$H_n = \left(\frac{1}{i + j - 1}\right)_{i,j=1}^{n},$$

which belongs to the class of Cauchy matrices ($t_i = i$, $s_j = 1 - j$), we have for $n = 10$, $c(F, a) \leqq 740$, while $c(G, H_{10}) \simeq 10^{12}$ and $k(G, H_{10}) \simeq 10^{16}$. It is also shown that one structured algorithm for computing the inverse of a Cauchy matrix is stable and produces accurate results when general matrix inversion algorithms fail.

Example 2. Let S be the set of all nonsingular Toeplitz matrices $A = (a_{i-j})_{i,j=0}^{n}$. Then $Ha = \sum_{k=-n}^{n} a_k Z_k$, where Z_k is an $(n + 1) \times (n + 1)$ matrix with 1's in positions of a_k's in A and zeros elsewhere. Let G be the map of matrix inversion. Then it is clear that for $k = -n, \ldots, n$

$$\frac{\partial F_{ij}}{\partial a_k} = \frac{\partial G_{ij}}{\partial Z_k}, \qquad i, j = 0, \ldots, n,$$

where $\partial G_{ij}/\partial Z_k$ is the directional derivative of the (i, j)th element of A^{-1} in the direction of Z_k at A. Thus

$$\frac{\partial F_{ij}}{\partial a_k} = r_i Z_k c_j,$$

where $c_0, \ldots, c_n$ are the columns of A^{-1}, and $r_0, \ldots, r_n$ are the rows of A^{-1}. The numerator in the formula for the mixed condition number $m(F, a)$ is given therefore by

$$\|F'(a)D_a\|_\infty = \max_{i,j=0,\ldots,n} \sum_{k=-n}^{n} |a_k|\,|r_i Z_k c_j|.$$

If $A = (a_{|i-j|})_{i,j=0}^{n}$ is symmetric, then

$$\|F'(a)D_a\|_\infty = \max_{i,j=0,\ldots,n} \sum_{k=0}^{n} |a_k|\,|r_i Z_k c_j|, \tag{2.11}$$

where Z_k is now a symmetric matrix with 1's in positions of a_k's in A and zeros elsewhere. Similar formulas can be obtained for other classes of matrices defined in Van Dooren [20] as pattern matrices, e.g., Hankel, circulant, and Sylvester matrices. Detailed analysis of structured condition numbers of Toeplitz and other structured matrices will be given in a forthcoming paper by the authors. Here we will only mention that, using (2.11), it is not hard to show that for a positive definite Toeplitz matrix A the following inequalities hold:

$$\frac{1}{(n+1)^2} m(G, A) \leqq m(F, a) \leqq m(G, A).$$

We also remark that using the Gohberg–Semencul formula for A^{-1} and the fast Fourier transform we can compute $r_i Z_k c_j$, $i, j, k = 0, \ldots, n$ in O $(n^3 \log n)$ flops, and for any fixed i or j (which corresponds to the structured condition number of the ith row or jth column of A^{-1}, respectively) in O $(n^2 \log n)$ flops. A different technique for computing structured condition numbers for linear systems of equations with Toeplitz and other structured coefficient matrices can be found in Higham and Higham [12].

Let $\lambda \neq 0$ be a simple eigenvalue of a Toeplitz matrix A and let $G : A \rightarrow \lambda$. Then it is well known (see Wilkinson [21]) that

$$\frac{\partial G}{\partial Z_k} = \frac{y^T Z_k x}{y^T x}, \qquad k = 1, \ldots, n,$$

where x and y are the properly normalized right and left eigenvectors corresponding to λ. Thus the mixed structured condition number in this case is given by

$$m(F, a) = \frac{1}{|\lambda| \, |y^T x|} \sum_{k=-n}^{n} |a_k| \, |y^T Z_k x|.$$

In the next section we analyze structured condition numbers of Vandermonde matrices and present a new structured algorithm for their inversion.

3. Applications to Vandermonde matrices. In this section, we consider computational problems with Vandermonde matrices $A = \{a_j^i\}_{i,j=0}^n$, where $a_i \neq a_j$ for $i \neq j$. The main emphasis here is on the case when $a_0, \ldots, a_n$ have arbitrary signs. The case of constant signs is treated in detail by Higham [10].

3.1. Matrix inversion. It is well known (see, for example, Knuth [13, p. 36]) that A^{-1} exists and

$$A_{ij}^{-1} = (-1)^j \frac{p_{ij}(a)}{q_i(a)}, \qquad i, j = 0, \ldots, n, \tag{3.1}$$

where

$$p_{ij}(a) = \sum_{\substack{0 \leqq m_0 < \cdots < m_{n-j-1} \leqq n \\ m_0, \ldots, m_{n-j-1} \neq i}} a_{m_0} \cdots a_{m_{n-j-1}}, \qquad i, j = 0, \ldots, n, \tag{3.2}$$

and

$$q_i(a) = \prod_{\substack{m=0 \\ m \neq i}}^{n} (a_m - a_i), \qquad i = 0, \ldots, n. \tag{3.3}$$

In the notation of § 2.5 G is the matrix inversion map, S is the subset of D_G consisting of all Vandermonde matrices, H is defined on all $n + 1$-tuples of pairwise different numbers

$$H : (a_0, \ldots, a_n) \rightarrow (a_j^i)_{i,j=0}^n,$$

and F is given by

$$F : (a_0, \dots, a_n) \to \left((-1)^j \frac{p_{ij}(a)}{q_i(a)} \right)_{i,j=0}^n .$$

The componentwise structured condition number $c(F, a)$ is not defined for a general $\mathbf{a}$, because Fa can have zero components. For example, $p_{n,n-1}(a) = a_1 + a_2 + \cdots + a_n$. Therefore, we will estimate the mixed structured condition number $m(R_i, a)$, $i = 0, \dots, n$, where R_i maps a into the ith row of A^{-1},

$$R_i : (a_0, \dots, a_n) \to \left((-1)^j \frac{p_{ij}(a)}{q_i(a)} \right)_{j=0}^n .$$

THEOREM 1. *Let R_i map $a = (a_0, \dots, a_n)$ into the ith row of the inverse of the Vandermonde matrix $A = (a_j^i)_{i,j=0}^n$. Then*

$$m(R_i, a) \leqq n^2 \max (n\Delta, n + \Delta), \tag{3.4}$$

where

$$\Delta = \max_{\substack{i,j=0,n \\ i \neq j}} \frac{|a_i|}{|a_i - a_j|} . \tag{3.5}$$

Proof. It is enough to show (3.4) for $i = 0$. For other rows the proof is exactly the same. Thus

$$R_0 a = q_0^{-1}(a)((-1)^j p_{0j}(a))_{j=0}^n .$$

We will estimate the ratio $\|R_0'(a)D_a\| / \|R_0 a\|$ in 1-norm first, and then will pass to the ∞-norm. We have

$$R_0'(a)D_a = \left\{ (-1)^j a_k \left[\frac{1}{q_0(a)} \frac{\partial}{\partial a_k} p_{0j}(a) + p_{0j}(a) \frac{\partial}{\partial a_k} \frac{1}{q_0(a)} \right] \right\}_{j,k=0}^n . \tag{3.6}$$

Since $(\partial/\partial a_0) p_{0j}(a) = 0$ for all $j = 0, \dots, n$, it follows that the first column of $R_0'(a)D_a$ is of the form

$$c_0 = \left((-1)^j a_0 p_{0j}(a) \frac{\partial}{\partial a_0} \left(\frac{1}{q_0(a)} \right) \right)_{j=0}^n .$$

Thus

$$\frac{\|c_0\|_1}{\|R_0 a\|_1} = \left| a_0 q_0(a) \frac{\partial}{\partial a_0} \left(\frac{1}{q_0(a)} \right) \right| = \left| a_0 \sum_{m=1}^n \frac{1}{a_m - a_0} \right| \leqq n\Delta.$$

For $k = 1, \dots, n$ the kth column of $R_0'(a)D_a$ is of the form $c_k = c_k^1 + c_k^2$, where

$$c_k^1 = q_0^{-1}(a) \left((-1)^j a_k \frac{\partial}{\partial a_k} p_{0j}(a) \right)_{j=0}^n$$

and

$$c_k^2 = \left((-1)^j a_k p_{0j}(a) \frac{\partial}{\partial a_k} \left(\frac{1}{q_0(a)} \right) \right)_{j=0}^n .$$

Thus

$$\frac{\|c_k^2\|_1}{\|R_0 a\|_1} = \left| a_k q_0(a) \frac{\partial}{\partial a_k}\left(\frac{1}{q_0(a)}\right)\right| = \frac{|a_k|}{|a_k - a_0|} \leqq \Delta.$$

Let us now consider c_k^1. It follows from (3.2) that

$$\begin{pmatrix} \dfrac{\partial}{\partial a_k} p_{00}(a) \\ \vdots \\ \dfrac{\partial}{\partial a_k} p_{0,n-1}(a) \end{pmatrix} = \begin{pmatrix} 1 & -a_k & a_k^2 & \cdots & (-1)^{n-1}(-a_k)^{n-1} \\ 0 & 1 & -a_k & & \vdots \\ \vdots & & & & a_k^2 \\ & & & 1 & -a_k \\ 0 & & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} p_{01}(a) \\ \vdots \\ p_{0n}(a) \end{pmatrix}.$$

This is easy to see by using the inverse of the above matrix, which is bidiagonal with unit diagonal and a_k above the diagonal. Hence, if $|a_k| \leqq 1$, then

$$\begin{aligned} \|c_k^1\|_1 &\leqq \frac{|a_k|}{|q_0(a)|} \sum_{m=0}^{n-1} |a_k|^m \|((-1)^j p_{0j}(a))_{j=1}^n\|_1 \\ &\leqq \frac{n}{|q_0(a)|} \|((-1)^j p_{0j}(a))_{j=0}^n\|_1. \end{aligned}$$

(Note that $p_{0n}(a) = 1$, and hence $(\partial/\partial a_k)p_{0n}(a) = 0$.) It also follows from (3.2) that

$$a_k \begin{pmatrix} \dfrac{\partial}{\partial a_k} p_{00}(a) \\ \vdots \\ \dfrac{\partial}{\partial a_k} p_{0n-1}(a) \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -a_k^{-1} & 1 & & \vdots \\ \vdots & & & 0 \\ (-1)^n - (-a_k)^{-n+1} & & -a_k^{-1} & 1 \end{pmatrix} \begin{pmatrix} p_{00}(a) \\ \vdots \\ p_{0n-1}(a) \end{pmatrix},$$

and hence, if $|a_k| \geqq 1$, then

$$\begin{aligned} \|c_k^1\|_1 &\leqq \frac{1}{|q_0(a)|} \sum_{m=0}^{n-1} |a^{-m}| \, \|((-1)^j p_{0j}(a))_{n=0}^{n-1}\| \\ &\leqq \frac{n}{|q_0(a)|} \|((-1)^j p_{0j}(a))_{j=0}^n\|_1. \end{aligned}$$

In any case, $\|c_k^1\|/\|R_0 a\|_1 \leqq n$ and

$$\frac{\|c_k\|_1}{\|R_0 a\|_1} \leqq \frac{\|c_k^1\|_1}{\|R_0 a\|_1} + \frac{\|c_k^2\|_1}{\|R_0 a\|_1} \leqq n + \Delta, \qquad k = 1, \dots, n.$$

Therefore,

$$\frac{\|R_0'(a)D_a\|_1}{\|R_0 a\|_1} = \max_{k=0,\dots,n} \frac{\|c_k\|_1}{\|R_0 a\|_1} \leqq \max(n\Delta, n + \Delta).$$

To complete the proof note that from (2.7)

$$m(R_0, a) = \frac{\|R_0'(a)D_a\|_\infty}{\|R_0 a\|_\infty} \leqq n^2 \frac{\|R_0'(a)D_a\|_1}{\|R_0 a\|_1}. \qquad \square$$

We remark that since the ith row of A^{-1} consists of coefficients of the elementary Lagrange polynomial $L_i(z)$ such that $L_i(a_i) = 1$ and $L_i(a_j) = 0$ for $j = 0, \dots, n, j \neq i$, it follows that (3.4) gives a bound of the condition number of computing coefficients of $L_i(z)$.

Theorem 1 gives an a priori bound on $m(R_0, a)$. The approximate value of $m(R_0, a)$ can be computed a posteriori via (3.6) using expressions for $(\partial/\partial a_k)q_0^{-1}(a)$ and $(\partial/\partial a_k)p_0(a)$ obtained in the proof of the theorem. Since the computation of $(\partial/\partial a_k)p_0(a)$, $k = 1, \ldots, n$, amounts to the solution of n linear systems of equations with a bidiagonal coefficient matrix, it follows that $m(R_0, a)$ can be computed in O (n^2) flops. A different technique for computing structured condition numbers of Vandermonde matrices is given in Higham [10]. The bound (3.4) shows that the structured condition number $m(R_0, a)$ can be much smaller than the usual condition number $k(A) = \|A\| \|A^{-1}\|$ as the latter can grow exponentially with n (see Gautschi and Inglese [7]). The following example shows that even for small values of n the difference between $m(R_0, a)$ and $m(G, A)$, where G maps A into the zeroth row of A^{-1}, can be very significant. Let $a_k = (100 + k)^{-1}$, $k = 0, \ldots, n$. Then for $n = 5$, $\Delta = 105$ and bound (3.4) is equal to 13,125, while $m(G, A) \approx 10^{11}$. For $n = 10$ we have $\Delta = 110$, and the bound of (3.4) is equal to $1.1 \cdot 10^5$, while $m(G, A) \approx 5 \cdot 10^{17}$. Numerical computation of $m(R_0, a)$ also shows that the bound of (3.4) differs from $m(R_0, a)$ by a factor that is approximately equal to n^2. This factor of n^2 comes from passing from 1-norm to ∞-norm in the last line of the proof of Theorem 1. Moreover, in a large number of numerical experiments (reported below in Example 1), we observed that $m(R_i, a)$ never exceeded 1.5 max $(n\Delta, n + \Delta)$ and that it was always greater than Δ. Therefore, we suggest using max $(n\Delta, n + \Delta)$ as an a priori approximate upper bound for $m(R_i, a)$, and Δ as an approximate lower bound.

The comparison of structured and general condition numbers gives additional justification to conclusions in Björck and Pereyra [4] and Higham [10] that structured algorithms for computation with Vandermonde matrices should be preferred to general purpose algorithms.

The natural question now is to find a stable structured algorithm for computing rows of the inverse of a Vandermonde matrix. Stability of the Björck–Pereyra algorithm for this problem in the case when $a_0, \ldots, a_n$ are of the same sign is established in Higham [10]. Another structured algorithm is given by Traub [19]. Let us describe these two algorithms.

ALGORITHM 1 (BJÖRCK-PEREYRA).

```
(i) $c_0^0 = 1,\ c_j^0 = 0 \quad (j = 1, \ldots, n)$.
    For $k = 0$ to $n - 1$
        $c_j^{(k+1)} = \dfrac{c_j^{(k)} - c_{j-1}^{(k)}}{a_j - a_{j-k-1}} \quad (j = n, n-1, \ldots, k+1)$,
        $c_j^{(k+1)} = c_j^{(k)} \quad (j = 0, \ldots, k)$.
(ii) $y_j^{(n)} = c_j^n \quad (j = 0, \ldots, n)$.
    For $k = n - 1$ to 0 step $-1$
        $y_j^{(k)} = y_j^{(k+1)} - a_k y_{j+1}^{(k+1)} \quad (j = k+1, \ldots, n-1)$,
        $y_j^k = y_j^{(k+1)} \quad (j = 0, \ldots, k, n)$.
(iii) $R_0 a = y^{(0)}$.
```

This is the algorithm for dual systems in Björck and Pereyra [4]. It computes $R_0 a$ in O (n^2) flops and A^{-1} in O (n^3) flops.

Traub [19] presents an O (n^2) algorithm for computing the whole A^{-1}. However, as indicated in Higham [10], this algorithm does not have the crucial property of the

Björck–Pereyra algorithm, namely, that there are no subtractions of numbers with the same sign in the case when $a_0, \dots, a_n$ have the same sign. The Traub algorithm can be easily modified to satisfy this property as follows. (Note that it is done on expense of making it an O (n^3) algorithm for computing the whole A^{-1}.)

ALGORITHM 2 (TRAUB).

(i) $c_0^0 = 1$

(ii) For $k = 0$ to $n - 1$

$$\begin{bmatrix} c_0^{k+1} \\ c_1^{k+1} \\ \vdots \\ c_{k+1}^{k+1} \end{bmatrix} = a_{k+1} \begin{bmatrix} c_0^k \\ \vdots \\ c_k^k \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ c_0^k \\ \vdots \\ c_k^k \end{bmatrix}.$$

(iii) $R_0 a = \dfrac{1}{q_0(a)} ((-1)^j c_j^n)_{j=0}^n.$

A numerical example at the end of this section demonstrates that neither of these two algorithms is stable for computing $R_0 a$ when **a** has entries with different signs. To derive a new algorithm we observe first that for $k = 0, \dots, n$,

$$\sum_{j=0}^{n} (-1)^j p_{kj}(a) z^j = \prod_{\substack{m=0 \\ m \neq k}}^{n} (a_m - z), \tag{3.7}$$

where z is any complex number. In particular, if $z = w_j = e^{2\pi i[j/(n+1)]}$, then

$$A^{-1} \begin{bmatrix} 1 \\ w_j \\ \vdots \\ w_j^n \end{bmatrix} = \begin{bmatrix} q_0^{-1}(a) \prod_{m=1}^{n} (a_m - w_j) \\ \vdots \\ q_n^{-1}(a) \prod_{m=0}^{n-1} (a_m - w_j) \end{bmatrix}.$$

Thus, if $\mathcal{F} = (w_j^k)_{k,j=0}^n$ is the discrete Fourier transform matrix, then

$$A^{-1}\mathcal{F} = D_q^{-1} S$$

where $D_q = \operatorname{diag}\{q_0(a), \dots, q_n(a)\}$ and

$$S_{kj} = \prod_{\substack{m=0 \\ m \neq k}}^{n} (a_m - w_j), \qquad k, j = 0, \dots, n.$$

Therefore,

$$A^{-1} = D_q^{-1} S \mathcal{F}^{-1}, \tag{3.8}$$

where $\mathcal{F}^{-1} = [1/(n+1)](\bar{w}_j^k)_{j,k=0}^n$. The ith row of A^{-1} is of the form

$$R_i a = q_i^{-1}(a) \left[\prod_{\substack{m=0 \\ m \neq i}}^{n} (a_m - w_0), \dots, \prod_{\substack{m=0 \\ m \neq k}}^{n} (a_m - w_n) \right] \mathcal{F}^{-1}.$$

The following algorithm is for the zeroth row. The rest can be found similarly.

ALGORITHM 3.

```
(i) $q = \prod_{m=1}^{n} (a_m - a_0)$.
(ii) For $j = 0$ to $n$
        $z_j = \prod_{m=1}^{n} (a_m - w_j)$.
(iii) $R_0 a = q^{-1}[z_0, \dots, z_n]\mathscr{F}^{-1}$.
```

This is an O (n^2) algorithm for computing one row of A^{-1} and O (n^3) for the whole A^{-1}. We remark that the computation of q gives small forward error (see Wilkinson [21]). The computation of z_j's amounts to multiplication by scaled orthogonal matrices, because if

$$\alpha + i\beta = \prod_{m=1}^{n} (\alpha_m + i\beta_m), \quad \text{then} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \alpha_1 & -\beta_1 \\ \beta_1 & \alpha_1 \end{bmatrix} \cdots \begin{bmatrix} \alpha_{n-1} & -\beta_{n-1} \\ \beta_{n-1} & \alpha_{n-1} \end{bmatrix} \begin{bmatrix} \alpha_n \\ \beta_n \end{bmatrix}.$$

The final step in computing $R_0 a$ is the multiplication of a vector by a unitary matrix. Since multiplication by unitary matrices is considered harmless, we may conjecture that Algorithm 3 will give errors proportional to the structured condition number. This conjecture is supported by the following numerical experiment. (Note that the algorithm and the representation (3.8) for A^{-1} remain valid when $a_0, \dots, a_n$ are complex numbers.)

Experiment 1. To illustrate numerical properties of Algorithms 1, 2, and 3 we compare them on the following sets of Vandermonde matrices. Let $a_k = \cos[(2k-1)/2n]\pi$, $k = 1, \dots, n$, be the roots of the nth-degree Chebychev polynomial $T_n(x)$, and let a_0 be chosen randomly in $(-1, 1)$. Thus the first row of A^{-1}, y, consists of coefficients of $T_n(x)$ (after scaling by $T_n(a_0)$), which are computed as integers from the recursion $T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$. Let y_1, y_2, and y_3 be computed via Algorithms 1, 2, and 3, respectively, and let

$$\varepsilon_i = \frac{\|y_i - y\|_\infty}{\|y\|_\infty}, \qquad i = 1, 2, 3.$$

In Table 1 we present the worst errors ε_i over 20 runs, with a_0 chosen randomly in $(-1, 1)$ for each run. This is done for $n = 10, 80(10)$. Because of round-off errors, the entries in computed y_3 have nonzero imaginary parts, so that, in fact,

$$\varepsilon_3 = \frac{\|\mathrm{Re}\,(y_3) - y\|_\infty}{\|y\|_\infty}.$$

TABLE 1

n	ε_1	ε_2	ε_3
10	3.0×10^{-6}	3.0×10^{-6}	3.0×10^{-6}
20	1.6×10^{-5}	9.0×10^{-6}	2.5×10^{-6}
30	1.7×10^{-4}	9.0×10^{-5}	1.1×10^{-5}
40	4.0×10^{-3}	9.0×10^{-4}	0.9×10^{-6}
50	5.0×10^{-2}	1.8×10^{-2}	1.4×10^{-5}
60	1.7	0.1	5.6×10^{-6}
70	19	13	4.9×10^{-6}
80	2687	5221	3.0×10^{-6}

For $n > 80$, overflows are encountered. The experiment was carried out on a SUN 3/50 workstation in single precision ($u \simeq 10^{-7}$). In all cases, we computed $m(R_0, a)$ and found that $m(R_0, a) \leqq 5 \cdot 10^3$ always. Since the computer precision is approximately 10^{-7} we can conclude that Algorithms 1 and 2 (from Björck–Pereyra and Traub, respectively) are unstable on the class of all Vandermonde matrices. We also computed $m(R_i, a)$ for different i's and various choices of $a_0, \dots, a_n$, and found that in all cases

$$\Delta < m(R_i, a) < 1.5 \max(n\Delta, n + \Delta).$$

3.2. Solution of $Ay = b$. The representation (3.7) of the inverse of a Vandermonde matrix can be used for the solution of $Ay = b$. The computation of entries of S in (3.8) requires O (n^3) flops, but it can be reduced to O (n^2) flops as follows. If $a_j \neq w_k, j, k = 0, \dots, n$, then

$$S = CD_w,$$

where

$$D_w = \operatorname{diag}\left\{\prod_{m=0}^{n} (a_m - w_0), \dots, \prod_{m=0}^{n} (a_m - w_n)\right\}$$

and

$$C = \left(\frac{1}{a_k - w_j}\right)_{k,j=0}^{n}.$$

Otherwise, there always exists a point on the unit circle, $\alpha = e^{i\theta}$, such that

$$e^{i\theta} a_j \neq w_k, \qquad j, k = 0, \dots, n.$$

Then the solution of $Ay = b$ is also the solution of $D_\alpha Ay = D_\alpha b$, where $D_\alpha = \operatorname{diag}\{1, \alpha, \dots, \alpha^n\}$. Since $D_\alpha A$ is a Vandermonde matrix defined by $\alpha a_0, \dots, \alpha a_n$, it follows that

$$y = \frac{1}{\alpha^n} D_q^{-1} C_\alpha D_{\alpha w} \mathscr{F}^{-1} D_\alpha b, \tag{3.9}$$

where

$$C_\alpha = \left(\frac{1}{\alpha a_k - w_j}\right)_{k,j=0}^{n}$$

and

$$D_{\alpha w} = \operatorname{diag}\left\{\prod_{m=0}^{n} (\alpha a_m - w_0), \dots, \prod_{m=0}^{n} (\alpha a_m - w_n)\right\}.$$

The computation of y via (3.9) requires O (n^2) flops on a sequential computer and O (n) flops on a single instruction multiple data (SIMD) parallel machine. We remark that since $|\alpha| = 1$, it is not hard to show, using techniques similar to those in the proof of Theorem 1, that multiplication of **a** by α does not affect the structured condition numbers, namely, that $m(R_i, a) = m(R_i, \alpha a), i = 0, \dots, n$. When $a_0, \dots, a_n$ are real numbers, then we can choose $\theta = \pi/(n + 1)$, in which case

$$|\alpha a_k - w_j| \geqq \sin\theta, \qquad k, j = 0, \dots, k.$$

When $a_0, \dots, a_n$ are complex numbers, the choice of α is more complicated and requires additional study.

TABLE 2

n	ε_1	ε_2
10	9.3×10^{-6}	1.4×10^{-6}
20	2.7×10^{-2}	5.0×10^{-6}
30	3	4.0×10^{-5}
40	22	1.2×10^{-4}
50	400	3.6×10^{-2}
60	312	1.4×10^{-3}
70	898	2.0×10^{-4}

Next, we present results of a numerical experiment in which the solution (3.9) compares favorably with the Björck–Pereyra [4] algorithm for primal systems.

Experiment 2. The points $a_0, \dots, a_n$ are chosen randomly in $(-1, 1)$, which implies that we can take $\alpha = 1$ in (3.9). The solution of $Ay = e_n$ is computed via the formula

$$y = \left(\frac{1}{q_i(a)}\right)_{i=0}^{n}$$

in double precision. The solutions y_1 and y_2 are computed in single precision via the Björck–Pereyra algorithm for primal systems and (3.9), respectively, and

$$\varepsilon_i = \frac{\|y_i - y\|_\infty}{\|y\|_\infty}, \qquad i = 1, 2.$$

In Table 2 we present the worst errors ε_1 and ε_2 over 20 runs with $(a_0, \dots, a_n)$ chosen at random in $(-1, 1)$ for each run. This is done for $n = 10, 70(10)$. Overflows occur for $n > 70$. The experiment was carried out on a SUN 3/50 workstation.

The same experiment has been carried out using (3.9) with $\alpha = i\pi/e^{n+1}$, with very similar results. Similar results for this α have also been obtained for cases when $a_0, \dots,$ a_n have been chosen at random in $(-1, 1)$ and then a_0 has been set to 1 for even n and to 1 or -1 for odd n.

REFERENCES

[1] M. ARIOLI, J. W. DEMMEL, AND I. S. DUFF, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.

[2] I. BABUŠKA, *Numerical stability in problems of linear algebra*, SIAM J. Numer. Anal., 9 (1972), pp. 53–77.

[3] G. R. BELITSKII AND Y. I. LYUBICH, *Matrix Norms and Their Applications*, OT 36, Birkhäuser, Basel, 1988.

[4] A. BJÖRCK AND V. PEREYRA, *Solution of Vandermonde systems of equations*, Math. Comp., 24 (1970), pp. 893–903.

[5] J. R. BUNCH, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl., 88/89 (1987), pp. 49–66.

[6] J. DEMMEL AND W. KAHAN, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 873–912.

[7] W. GAUTSCHI AND G. INGLESE, *Lower bounds for the condition number of Vandermonde matrices*, Numer. Math., 52 (1988), pp. 241–250.

[8] I. GOHBERG AND I. KOLTRACHT, *On the inversion of Cauchy matrices*, in Proc. of Internat. Symposium MTNS-89, Vol. III, Birkhäuser, Basel, pp. 381–392.

[9] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[10] N. J. HIGHAM, *Error analysis of the Björck–Pereyra algorithms for solving Vandermonde systems*, Numer. Math., 50 (1987a), pp. 613–632.

[11] N. J. HIGHAM, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987b), pp. 575–596.

[12] D. J. HIGHAM AND N. J. HIGHAM, *Backward error and condition of structured linear systems*, University of Manchester, England/UMIST, Numerical Analysis Report No. 192, 1990; SIAM J. Matrix Anal. Appl., 13 (1992), pp. 162–175.

[13] D. E. KNUTH, *The Art of Computer Programming*, Addison-Wesley, Reading, MA, 1968.

[14] Y. I. LYUBICH, *Conditioning of general computational problems*, Dokl. Akad. Sc. USSR, 171 (1966), pp. 791–793. (In Russian.)

[15] J. R. RICE, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.

[16] J. ROHN, *New condition numbers for matrices and linear systems*, Computing, 41 (1989), pp. 167–169.

[17] R. D. SKEEL, *Scaling for numerical stability in Gaussian elimination*, J. Assoc. Comput. Mach., 26 (1979), pp. 494–526.

[18] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, Springer-Verlag, Berlin, New York, 1980.

[19] J. F. TRAUB, *Associated polynomials and uniform methods for the solution of linear problems*, SIAM Rev., 8 (1966), pp. 277–301.

[20] P. M. VAN DOOREN, *Structured Linear Algebra Problems in Digital Signal Processing*, Proceedings of NATO ASI, Leuven, 1988; Series F, Springer-Verlag, Berlin, New York, 1990.

[21] J. H. WILKINSON, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 705-711, July 1993

008

AN IMPROVEMENT OF HADAMARD'S INEQUALITY FOR TOTALLY NONNEGATIVE MATRICES*

ZHANG XIAODONG† AND YANG SHANGJUN†

Abstract. In this paper an improvement is obtained of Hadamard's inequality for totally nonnegative matrices: If $A = (a_{ij})$ is an $n \times n$ totally nonnegative matrix with $a_{11}a_{22} \cdots a_{nn} \neq 0$, then

$$\det A \leqq \min \left\{ \prod_{i=1}^{n} a_{ii} - \max_{1 \neq \sigma \in S_n} \left(\prod_{i=1}^{n} a_{i\sigma(i)} a_{\sigma(i)i} \right)^{1/2},\ \min a_{kk} \prod_{\substack{i=1 \\ i \neq k}}^{n} \left(a_{ii} - \frac{a_{ki}a_{ik}}{a_{kk}} \right) \right\},$$

and the inequality is a strict one when A is totally positive and $n \geqq 3$. Then we prove that for an $n \times n$ totally nonnegative matrix A, $\det A = a_{11}a_{22} \cdots a_{nn}$ if and only if each diagonal product, except the main diagonal product, of A has a zero entry on it.

Key words. totally nonnegative, totally positive, Hadamard's inequality. Sylvester's identity

AMS subject classification. 15A

1. Introduction and preliminaries.

The following theorem is a well-known classical result (cf. [4]).

HADAMARD'S THEOREM. *If $A = (a_{ij})$ is an $n \times n$ positive semidefinite Hermitian matrix, then*

$$\det A \leqq \prod_{i=1}^{n} a_{ii}.$$

Furthermore, equality holds if and only if A is diagonal or A has a zero row or column.

It is known (cf. [5]) that Hadamard's inequality holds for a totally nonnegative matrix A. More interesting results related to Hadamard's inequality can be found in [2], [3], [6], etc.

In this paper we shall obtain some inequalities stronger than Hadamard's inequality for totally nonnegative and totally positive matrices, and investigate the necessary and sufficient conditions for equality to hold.

A real $n \times n$ matrix A is called *totally nonnegative* (*positive*), if all subdeterminants of A of all orders are nonnegative (positive). Following [1] we use $Q_{k,n}$ to denote all $\binom{n}{k}$ increasing sequences $w = (w_1, \ldots, w_n)$, $1 \leqq w_1 < w_2 < \cdots < w_k \leqq n$. If $A = (a_{ij})$ is an $n \times n$ matrix, $\alpha, \beta \in Q_{k,n}$, then $A[\alpha|\beta]$ denotes the $k \times k$ submatrix of A whose (i, j) entry is $a_{\alpha_i\beta_i}$. For a principal submatrix $A[\alpha|\alpha]$ we use the more abbreviated notation $A[\alpha]$.

We need the following results, which can be found in [5].

THEOREM 1 (Fischer's inequality). *If A is an $n \times n$ totally nonnegative matrix and $1 \leqq k \leqq n-1$, then*

$$\det A \leqq \det(A[1, \ldots, k]) \det(A[k+1, \ldots, n]).$$

THEOREM 2 (Sylvester's identity). *For an $n \times n$ matrix $B = (b_{ij})$ and $1 \leqq k \leqq n-1$ define the $(n-k) \times (n-k)$ matrix $D = (d_{ij})$ by*

$$d_{ij} = \det(B[1, \ldots, k, k+i|1, \ldots, k, k+j]), \qquad 1 \leqq i, j \leqq n-k;$$

* Received by the editors April 8, 1991; accepted for publication (in revised form) December 2, 1991.

† Department of Mathematics, Anhui University, Hefei, Anhui 230039, China.

then D is totally nonnegative (positive) if B is totally nonnegative (positive), and

$$\det D = (\det B[1, \ldots, k])^{n-k-1} \det B.$$

2. Improvement of Hadamard's inequality. In the sequel, we use notation $\langle n\rangle$ for $\{1, \ldots, n\}$, S_n for the symmetric group on $\langle n\rangle$, and 1 for the unit of S_n. To avoid triviality we always assume $n > 1$.

THEOREM 3. *If A is an $n \times n$ totally nonnegative matrix with some $a_{kk} > 0$, then*

$$(1) \qquad \det A \leqq a_{kk} \prod_{\substack{i=1 \\ i \neq k}}^{n} \left(a_{ii} - \frac{a_{ki}a_{ik}}{a_{kk}}\right).$$

Furthermore, (1) is a strict inequality when A is totally positive and $n \geqq 3$.

Proof. Consider the $(n-1) \times (n-1)$ matrix $D = (d_{ij})$ defined by

$$d_{ij} = \det(A[\alpha_{ki} | \alpha_{kj}]) \qquad i, j \in \langle n\rangle - \{k\},$$

where $\alpha_{kh} \in Q_{2,n}$ consists of k and h. Since A is totally nonnegative (positive), D is totally nonnegative (positive). Let E_{ik} denote the $n \times n$ (0, 1) matrix whose only 1 is at position (i, k), and

$$M_{ik} = I_n - \frac{a_{ik}}{a_{kk}} E_{ik}, \qquad i \in \langle n\rangle - \{k\}.$$

Then $\det M_{ik} = 1$ and $M_{1k} \cdots M_{k-1,k}M_{k+1,k} \cdots M_{nk}A = \bar{W} = (w_{ij})$ has its entries as follows:

$$w_{ij} = a_{ij} - \frac{a_{kj}a_{ik}}{a_{kk}}, \qquad i \in \langle n\rangle - \{k\}$$

$$w_{kj} = a_{kj}, \qquad j \in \langle n\rangle.$$

It is easily seen that

$$w_{ij} = \operatorname{sgn}(i-k) \operatorname{sgn}(j-k) \frac{d_{ij}}{a_{kk}} \quad \text{for } i, j \in \langle n\rangle - \{k\},$$

where $\operatorname{sgn}(x) = 1$ if $x > 0$, and -1 if $x < 0$. Let

$$J_h = I_n - 2E_{hh}, \quad \text{then } J_1 \cdots J_{k-1} W J_1 \cdots J_{k-1} = \frac{1}{a_{kk}} D.$$

Therefore,

$$(2) \qquad \begin{aligned} \det A &= a_{kk} \det(W[\langle n\rangle - \{k\}]) \\ &= a_{kk}(a_{kk})^{-(n-1)} \det D = (a_{kk})^{-n+2} \det D. \end{aligned}$$

In general, for any two sequences α, β in $Q_{h,n}$, so that $2 \leqq h \leqq n$ and $k \in \alpha \cap \beta$. We can similarly prove

$$\det(A[\alpha|\beta]) = (a_{kk})^{-h+2} \det(D[\alpha - \{k\} | \beta - \{k\}]),$$

whence D is totally nonnegative (positive) when A is totally nonnegative (positive). Now apply Hadamard's inequality for $\det D$ in (2) to obtain

$$\det A \leqq (a_{kk})^{-n+2} \prod_{\substack{i=1 \\ i \neq k}}^{n} d_{ii} = a_{kk} \prod_{\substack{i=1 \\ i \neq k}}^{n} \left(a_{ii} - \frac{a_{ki}a_{ik}}{a_{kk}}\right).$$

If A is totally positive and $n \geqq 3$ then D is a totally positive matrix of order at least 2, whence $\det D < \prod_{\substack{i=1\\ i\neq k}}^{n} d_{ii}$, and (1) is a strict inequality.

LEMMA 1. *If $A = (a_{ij})$ is an $n \times n$ totally nonnegative matrix, then*

$$\prod_{i=1}^{n} a_{ii} \geqq \left[\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i}\right]^{1/2} \tag{3}$$

for each $\sigma \in S_n$, where in the sequel, $(x)^{1/2}$ denotes the arithmetic square root of $x \geqq 0$. Furthermore, if A is totally positive, then (3) *is a strict inequality for any $1 \neq \sigma \in S_n$.*

Proof. Since each 2×2 principal minor of the totally nonnegative matrix A is nonnegative, we have

$$a_{ii}a_{\sigma(i)\sigma(i)} \geqq a_{i\sigma(i)}a_{\sigma(i)i}, \qquad i = 1, 2, \ldots, n. \tag{4}$$

Multiplying the n inequalities in (4) yields

$$\left(\prod_{i=1}^{n} a_{ii}\right)^2 \geqq \prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i},$$

whence inequality (3) follows.

If A is totally positive, then all the inequalities in (4) are strict. In this case, inequality (3) is strict.

LEMMA 2. *If A is an $n \times n$ totally nonnegative matrix, then*

$$\det A \leqq (a_{11}a_{nn} - a_{1n}a_{n1}) \prod_{i=2}^{n-1} a_{ii}, \tag{5}$$

and the inequality is strict when A is totally positive and $n \geqq 3$.

Proof. If $\det(A[1, \ldots, n-2]) = 0$, then $\det A = \det(A[1, \ldots, n-2]) \det(A \times [n-1, n]) = 0$ and (5) clearly holds. Thus we can assume $\det(A[1, \ldots, n-2]) > 0$.

We prove (5) by induction on n. Since (5) holds with equality when $n = 2$, we consider $n > 2$ and assume (5) holds for all totally nonnegative matrices of orders $m < n$. Consider the 2×2 matrix $D = (d_{ij})$, where

$$d_{ij} = \det(A[1, \ldots, n-2, n-2+i \,|\, 1, \ldots, n-2, n-2+j]), \quad 1 \leqq i, j \leqq 2.$$

Then D is totally nonnegative by Theorem 2;

$$d_{11} = \det(A[1, \ldots, n+1]) \leqq a_{n-1,n-1} \det(A[1, \ldots, n-2]) \tag{6}$$

by Fischer's inequality; and

$$d_{22} = \det(A[1, \ldots, n-2, n]) \leqq (a_{11}a_{nn} - a_{1n}a_{n1}) \prod_{i=2}^{n-2} a_{ii} \tag{7}$$

by the induction hypothesis, since $A[1, \ldots, n-2, n]$ is totally nonnegative. Now it follows from Sylvester's identity that

$$\det A = \frac{\det D}{\det(A[1, \ldots, n-2])} \leqq \frac{d_{11}d_{22}}{\det(A[1, \ldots, n-2])}$$

by Hadamard's inequality for D. Finally, use (6) and (7) to yield the desired result (5).

Observe that when A is totally positive, so is D, whence

$$\det D = d_{11}d_{22} - d_{12}d_{21} < d_{11}d_{22}.$$

In this case, inequality (5) is strict.

THEOREM 4. *If A is an $n \times n$ totally nonnegative matrix, then*

$$\det A \leqq \prod_{i=1}^{n} a_{ii} - \left[\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i}\right]^{1/2} \tag{8}$$

for each $\sigma \neq 1$ in S_n. Furthermore, (8) is a strict inequality when A is totally positive and $n \geqq 3$.

Proof. Since (8) becomes Hadamard's inequality if $\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i} = 0$, we assume

$$\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i} > 0. \tag{9}$$

First, we assert that for any $1 \neq \sigma \in S_n$ satisfying (9), there exists $k \neq \sigma(k)$ such that

$$\left(\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i}\right)^{1/2} \leqq a_{k\sigma(k)}a_{\sigma(k)k} \prod_{\substack{i=1\\ i\neq k,\sigma(k)}}^{n} a_{ii}. \tag{10}$$

If it is not the case, then

$$\left(\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i}\right)^{1/2} > a_{k\sigma(k)}a_{\sigma(k)k} \prod_{\substack{i=1\\ i\neq k,\sigma(k)}}^{n} a_{ii} \tag{11}$$

for any $k \in T = \{i \mid \sigma(i) \neq i,\ 1 \leqq i \leqq n\}$. Let $S = \{1, \ldots, n\} - T$, the complementary set of T. Multiplying all the possible inequalities in (11) yields

$$\left[\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i}\right]^{|T|/2} > \prod_{k\in T}\left(a_{k\sigma(k)}a_{\sigma(k)k} \prod_{i=1}^{n} a_{ii}\right),$$

where $|T|$ denotes the cardinal number of T. Now it follows that

$$\left(\prod_{i\in T} a_{i\sigma(i)}a_{\sigma(i)i}\right)^{|T|/2}\left(\prod_{i\in S} a_{ii}\right)^{|T|} > \prod_{i\in T} a_{i\sigma(i)}a_{\sigma(i)i}\left(\prod_{i\in S} a_{ii}\right)^{|T|}\left(\prod_{i\in T} a_{ii}\right)^{|T|-2}. \tag{12}$$

If $\prod_{i\in S} a_{ii} = 0$, then $\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i} = 0$ by Lemma 1, and since this contradicts our assumption (9), we must have $\prod_{i\in S} a_{ii} > 0$. Cancelling this positive factor from inequality (12) yields

$$\left(\prod_{i\in T} a_{i\sigma(i)}a_{\sigma(i)i}\right)^{(|T|-2)/2} > \left(\prod_{i\in T} a_{ii}\right)^{|T|-2},$$

whence

$$\left(\prod_{i\in T} a_{i\sigma(i)}a_{\sigma(i)i}\right)^{1/2} > \prod_{i\in T} a_{ii}. \tag{13}$$

On the other hand, since the principal submatrix $A[T]$ of A is totally nonnegative, applying Lemma 1 to $A[T]$ yields

$$\left(\prod_{i\in T} a_{i\sigma(i)}a_{\sigma(i)i}\right)^{1/2} \leqq \prod_{i\in T} a_{ii},$$

which is a contradiction to (13). Therefore, (10) holds.

Without loss of generality, suppose $k < \sigma(k) = h$ satisfies (10). Then by Fischer's inequality, Hadamard's inequality, and Lemma 2, we have

$$\begin{aligned} \det A &\leqq \det A[1, \dots, k-1] \det A[k, \dots, h] \det A[h+1, \dots, n] \\ &\leqq \left(\prod_{i=1}^{k-1} a_{ii}\right)\left(\prod_{i=h+1}^{n} a_{ii}\right) \det A[k, \dots, h], \end{aligned} \tag{14}$$

and

$$\det A[k, \dots, h] \leqq (a_{kk}a_{hh} - a_{kh}a_{hk}) \prod_{i=k+1}^{h-1} a_{ii}; \tag{15}$$

thus

$$\begin{aligned} \det A &\leqq (a_{kk}a_{hh} - a_{kh}a_{hk}) \prod_{i \neq k,h} a_{ii} \\ &= \prod_{i=1}^{n} a_{ii} - a_{k\sigma(k)}a_{\sigma(k)k} \prod_{i \neq k,\sigma(k)} a_{ii} \leqq \prod_{i=1}^{n} a_{ii} - \left(\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i}\right)^{1/2} \end{aligned} \tag{16}$$

by (10).

If A is totally positive, then each $a_{ii} > 0$. When $|T| \geqq 3$, instead of (10), we similarly prove

$$\left(\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i}\right)^{1/2} < a_{k\sigma(k)}a_{\sigma(k)k} \prod_{i \neq k,\sigma(k)} a_{ii}$$

by Lemma 2, whence inequality (16) is strict. In this case (8) is also strict. When $|T| = 2$ and $n \geqq 3$, we have $\sigma(k) = h$, $\sigma(h) = k$, and $\sigma(i) = i$ for $i \neq k, \sigma(k)$. Assume, without loss of generality, that $k < h$. If $h > k + 1$, then (15) is strict by Lemma 2, whence (8) is strict.

When $h = k + 1$, consider the 3×3 totally positive matrix $A[p, p+1, p+2]$, where $1 \leqq p \leqq k < p + 2 \leqq n$, and define the 2×2 positive matrix $D = (d_{ij})$ the same way as in the proof of Theorem 3, by

$$d_{ij} = \det(A[\alpha_{ki} | \alpha_{kj}]), \qquad i, j \in \{p, p+1, p+2\} - \{k\},$$

where

$$\alpha_{ki} = (k, i) \quad \text{if } k < i; \; \alpha_{ki} = (i, k) \quad \text{if } k > i.$$

Then D is totally positive and $\det(A[p, p+1, p+2]) = (\det D)/a_{kk}$, by equality (2). Furthermore,

$$\det D < d_{11}d_{22} < (a_{kk}a_{hh} - a_{kh}a_{hk})a_{kk}a_{rr},$$

where r is $k - 1$ or $h + 1$, and

$$\det(A[p, p+1, p+2]) < a_{rr}(a_{kk}a_{hh} - a_{kh}a_{hk}).$$

Finally,

$$\begin{aligned} \det A &= \left(\prod_{i=1}^{p-1} a_{ii}\right)\left(\prod_{i=p+3}^{n} a_{ii}\right) \det(A[p, p+1, p+2]) \\ &< (a_{kk}a_{hh} - a_{kh}a_{hk}) \prod_{\substack{i \neq k,\sigma(k) \\ i=1}}^{n} a_{ii} = \prod_{i=1}^{n} a_{ii} - \left(\prod_{i=1}^{n} a_{i\sigma(i)}a_{\sigma(i)i}\right)^{1/2}. \end{aligned}$$

Combining Theorems 3 and 4 yields our main result, which is Theorem 5.

THEOREM 5. *If A is an $n \times n$ totally nonnegative matrix with $a_{11} \cdots a_{nn} > 0$, then*

$$\det A \leqq \min \left\{ \prod_{i=1}^{n} a_{ii} - \max_{1 \neq \sigma \in S_n} \left(\prod_{i=1}^{n} a_{i\sigma(i)} a_{\sigma(i)i} \right)^{1/2}, \min_{k \in \langle n \rangle} a_{kk} \prod_{k \neq i = 1}^{n} \left(a_{ii} - \frac{a_{ki} a_{ik}}{a_{kk}} \right) \right\},$$

and this inequality is a strict one when A is totally positive and $n \geqq 3$.

3. When does equality in Hadamard's inequality hold? The last half of Hadamard's Theorem, stated in § 1, is no longer true for totally nonnegative matrices. As a counterexample, observe that the totally nonnegative matrix

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

satisfies $\det A = a_{11} a_{22}$, and it is neither diagonal nor does it have a zero row or column. But when does equality in Hadamard's inequality hold? The next theorem offers a complete answer to this question.

THEOREM 6. *Let A be an $n \times n$ totally nonnegative matrix, then*

$$\det A = a_{11} \cdots a_{nn} \tag{17}$$

if and only if each diagonal product, except the main diagonal, of A has a zero on it.

Proof. "If" is immediate. "Only if": Assume (17) holds for the totally nonnegative matrix A.

Case 1. $a_{11} \cdots a_{nn} > 0$. We prove it by induction on n. It is trivially true when $n = 2$. Suppose that any diagonal product, except possibly the main one, of each $(n-1) \times (n-1)$ totally nonnegative matrix, which satisfies (17), has a zero on it. We proceed to show that $n \times n$ matrix A shares the same property. Since $\det A = a_{11} \cdots a_{nn} > 0$, we have

$$a_{ki} a_{ik} = 0 \quad \text{for any } i \neq k \text{ in } \langle n \rangle \tag{18}$$

by (8). Assume, without loss of generality, that $a_{21} = 0$. Since

$$0 \leqq \det(A[2, j \mid 1, 2]) = -a_{22} a_{j1}, \qquad j = 3, \ldots, n.$$

we have

$$a_{31} = \cdots = a_{n1} = 0 \quad \text{and} \quad a_{11} \det(A[2, \ldots, n]) = \det A = a_{11} a_{22} \cdots a_{nn},$$

whence $\det(A[2, \ldots, n]) = a_{22} \cdots a_{nn}$. Thus each diagonal product except the main one of $A[2, \ldots, n]$ has a zero by the induction hypothesis. So each diagonal product of A has a zero.

Case 2. $a_{kk} = 0$ for some $k \in \langle n \rangle$. Assume, to the contrary, that there exists some $1 \neq \sigma \in S_n$, so that

$$a_{1\sigma(1)} \cdots a_{n\sigma(n)} \neq 0, \tag{19}$$

with $\sigma(u) = k$ and $\sigma(k) = v$; then, since

$$a_{kk} = 0 \quad \text{and} \quad \det(A[\alpha_{ku} \mid \alpha_{kv}]) \geqq 0,$$

we have

$$v > k \quad \text{if } u < k; \; v < k \quad \text{if } u > k.$$

Assume, without loss of generality, $u < k$, and hence $v > k$, since $a_{kk} = 0$, $a_{uk} > 0$, and $a_{kv} > 0$. Thus we have $a_{k1} = \cdots = a_{k,k-1} = 0 = a_{k+1,k} = \cdots = a_{nk}$ by the similar argument, and furthermore,

$$a_{ij} = 0 \quad \text{for any } i \geqq k, j \leqq k,$$

or $A[k, \ldots, n | 1, \ldots, k] = 0$, a contradiction to (19).

Observe that (17) holds trivially for any $n \times n$ totally nonnegative matrix A with $a_{11} \cdots a_{nn} = 0$. For the remaining case, we have the following theorem.

THEOREM 7. *For each $n \times n$ totally nonnegative matrix A with $a_{11} \cdots a_{nn} > 0$, A satisfies* (17) *if and only if*

$$(20) \qquad a_{i,i+1}a_{i+1,i} = 0 \quad \text{for } i = 1, \ldots, n-1.$$

Proof. Since (17) ⇒ (18) ⇒ (20), we only need to show that (20) ⇒ (17). We prove it by induction on n. It is obviously true when $n = 2$. Let $n > 2$ and assume that it is proved for $n - 1$. Since $a_{12}a_{21} = 0$, we have $a_{12} = 0$ or $a_{21} = 0$. When $a_{21} = 0$, we have shown in the proof of Theorem 6 that

$$\det A = a_{11} \det (A[2, \ldots, n]),$$

$$\det (A[2, \ldots, n]) = a_{22} \cdots a_{nn}.$$

Thus $a_{i,i+1}a_{i+1,i} = 0$ for $i = 2, \ldots, n-1$, by the induction hypothesis, whence $\det A = a_{11} \cdots a_{nn}$, i.e., (17) holds. We can similarly prove it when $a_{12} = 0$.

COROLLARY 1. *If an $n \times n$ totally nonnegative matrix A with $a_{11} \cdots a_{nn} > 0$ satisfies "$a_{ij} = 0 \Rightarrow a_{ji} = 0$ for any $i, j \in \langle n \rangle$," then A satisfies* (17) *if and only if A is diagonal.*

REFERENCES

[1] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979, pp. 49–50.

[2] G. M. ENGEL AND H. SCHNEIDER, *The Hadamard-Fischer inequality for a class of matrices defined by eigenvalue monotonicity*, Linear and Multilinear Algebra, 4 (1976), pp. 155–176.

[3] C. R. JOHNSON AND W. W. BARRETT, *Spanning-tree extensions of the Hadamard-Fischer inequalities*, Linear Algebra Appl., 66 (1985), pp. 177–193.

[4] H. MINC AND M. MARCUS, *A Survey of Matrix Theory and Matrix Inequalities*, Pridle, Weber & Schmidt, New York, 1964.

[5] H. MINC, *Nonnegative Matrices*, John Wiley & Sons, New York, 1988, pp. 152–154.

[6] G. MUHLBACH AND M. GASCA, *A generalization of Sylvester's identity on determinants and some applications*, Linear Algebra Appl., 66 (1985), pp. 221–234.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 712-720, July 1993

009

NECESSARY AND SUFFICIENT CONDITIONS FOR BOUNDED-INPUT/BOUNDED-STATE STABILITY OF DISCRETE-TIME, BILINEAR SYSTEMS*

G. O. CORRÊA†

Abstract. In this note, bounded-input/bounded state (BIBS) stability is studied for discrete-time multi-input, bilinear systems. The main result gives generic necessary and sufficient conditions for BIBS stability of such systems. These conditions comprise the existence of a single similarity transformation that takes to triangular form all the matrices defining a multi-input bilinear system, as well as constraints on their spectral radiuses.

Key words. bilinear systems, discrete-time systems, bounded-input/bounded-state stability

AMS subject classification. 93D05

NOTATION.
$\mathbb{N}$—the set of nonnegative integers
Z_+—the set of positive integers
$\mathbb{R}$—the set of real numbers
$\mathbb{R}_+$—the set of positive reals
$\mathbb{R}^n$—the set of real n-tuples
$\mathbb{R}^{n\times m}$—the set of real $n \times m$ matrices
$\mathbb{C}$—the set of complex numbers
$\mathbb{C}^n$—the set of complex n-tuples
$\mathbb{C}^{n\times m}$—the set of complex $n \times m$ matrices
$\|v\|$—Euclidean norm of $v \in \mathbb{C}^n$
$\|A\|$—corresponding induced operator norm (on $\mathbb{C}^n$) of $A \in \mathbb{C}^{n\times n}$
$\underline{\sigma}(A)$—smallest singular value of A
$\rho[A]$—spectral radius of A

1. Introduction. Bilinear systems have been considered in the survey papers by Bruni, DiPillo, and Koch [1] and Mohler and Kolodziej [2], which concentrated on continuous-time systems, whereas Goka, Tarn, and Zabersky [3] addressed some structural aspects of discrete-time, bilinear systems; the latter have also attracted some attention in the recent time-series and system identification literature (cf. Stensholt and Tjostheim [4], Subba Rao [5], and Fnaiech and Ljung [6]). More specifically, the stability of *stochastic*, discrete-time, bilinear systems has been addressed in Kubrusly and Costa [7], and the asymptotic stability of the origin of a (purely multiplicative) bilinear system with *constant*, *scalar* input has been investigated by Gounaridis-Minadis and Kalouptsidis [8] (again, for the discrete-time case).

In this note, bounded-input/bounded-state (BIBS) stability is studied for discrete-time, multi-input bilinear systems. Following the introduction of the required definition, a set of necessary conditions for BIBS stability is derived (Prop. 1). Then two lemmas are presented on the basis of which more explicit necessary conditions are obtained from Proposition 1. This leads to Theorem 1, which gives generic necessary and sufficient conditions for BIBS stability, with the proof of sufficiency being based on another lemma. Introducing an additional (generic) assumption on the bilinear systems considered, one of the conditions of Theorem 1 is made tighter (this is Theorem 2). Finally, necessary

* Received by the editors March 4, 1991; accepted for publication (in revised form) December 2, 1991.

† Laboratório Nacional de Computação Científica, Rua Lauro Müller 455, 22290, Rio de Janeiro, Brasil (goc@lncc.bitnet).

and sufficient conditions are presented for the asymptotic stability of the family of bilinear systems defined by bounded input sequences.

2. BIBS stability. Consider discrete-time, finite-dimensional, bilinear systems defined by

$$x(t+1) = \left[A_0 + \sum_{k=1}^{m} w_k(t)A_k\right]x(t) + Bw(t) + \bar{B}u(t), \qquad t \in \mathbb{N}, \tag{1}$$

where

$$x(t) \in \mathbb{R}^n, \quad A_i \in \mathbb{R}^{n\times n}, \quad B \in \mathbb{R}^{n\times m}, \quad \bar{B} \in \mathbb{R}^{n\times \bar{m}},$$
$$w(t)^T = [w_1(t) \cdots w_m(t)]^T \in \mathbb{R}^m, \qquad u(t) \in \mathbb{R}^{\bar{m}}.$$

State trajectories can be explicitly written as functions of initial conditions and input sequences as follows:

$$\begin{aligned} x(t) &= \{\mathscr{A}[w(t-1)] \cdots \mathscr{A}[w(0)]\}x(0) \\ &\quad + \sum_{l=1}^{t-1} \{\mathscr{A}[w(t-1)] \cdots \mathscr{A}[w(l)]\}[B \vdots \bar{B}]\begin{bmatrix} w(l-1) \\ u(l-1)\end{bmatrix} \\ &\quad + [B \vdots \bar{B}]\begin{bmatrix} w(t-1) \\ u(t-1)\end{bmatrix}, \end{aligned} \tag{2}$$

where

$$\mathscr{A}: \mathbb{R}^m \to \mathbb{R}^{n\times n},$$
$$w \to \mathscr{A}(w) = A_0 + \sum_{k=1}^{m} w_kA_k, \qquad w^T = [w_1 \cdots w_k].$$

DEFINITION 1 (input-state stability). The system $\Sigma = (\{A_k: k = 0, \ldots, m\}, B, \bar{B})$ given by (1) is said to be BIBS-stable from $x(0) = x \in \mathbb{R}^n$ if and only if for any pair of *bounded* sequences $(\{w(t)\}, \{u(t)\}: t \in \mathbb{N})$ the corresponding state-sequence $(\{x(t)\}: t \in \mathbb{N})$ is also bounded.

Necessary conditions for BIBS stability are now investigated. To cater to strictly bilinear systems (i.e., systems as in (1) without the term in u), all necessary conditions will be obtained with $u(t) \equiv 0$, or equivalently, with $\bar{B} = 0$.

PROPOSITION 1. *Let* $\Sigma = (\{A_l: l = 0, \ldots, m\}, B, 0)$ *denote a bilinear system as in* (1). *If either*

(i) Σ *is* BIBS-*stable from all* x *in a set of nonempty interior, or*

(ii) Σ *is* BIBS-*stable from some* $x \in S(\Sigma)$, *where*

$$S(\Sigma) \triangleq \{x \in \mathbb{R}^n: \text{rank}\,[M(x) + \mathscr{C}(A_0, B)] = n\}, \quad \textit{with}$$

$$\begin{aligned} M(x) \triangleq [&A_0^{n-1}A_1x \cdots A_0^{n-1}A_mx \vdots \cdots \vdots A_0^{n-i}A_1A_0^{i-1}x \cdots A_0^{n-i}A_mA_0^{i-1}x \vdots \\ &\cdots \vdots A_1A_0^{n-1}x \cdots \vdots A_mA_0^{n-1}x], \end{aligned}$$

$$\mathscr{C}(A_0, B) = [A_0^{n-1}B \vdots \cdots \vdots B]\},$$

then

$$\forall k \in Z_+ \quad \forall(\nu_1, \ldots, \nu_k) \in (\mathbb{R}^m)^k, \quad \rho[\mathscr{A}(\nu_k) \cdots \mathscr{A}(\nu_1)] \leqq 1,$$

where $\rho[A]$ *denotes the spectral radius of* A.

The proof of Proposition 1 is given in Corrêa [10]. Note that whenever (A_0, B) is controllable, $S(\Sigma)$ is *open* and *dense* in $\mathbb{R}^n$ and contains the origin. Note also that (i) is satisfied whenever Σ is BIBS-stable from some $\tilde{x}$ such that the set of states reachable from $\tilde{x}$ has nonempty interior.

The basic idea for the proof of Proposition 1 is to look at the state trajectories $\{x(t): t \geqq t_0,\ t \in \mathbb{N}\}$ obtained with periodic input sequences with period k, say, observing that such sequences are bounded if and only if the subsequences $\{x(t_0 + kq): q \in \mathbb{N}\}$ are bounded. These subsequences, in turn, are bounded only if the matrices $\{\mathscr{A}(\nu_k) \cdots \mathscr{A}(\nu_1)\}$ are power-bounded, the latter implying that the spectral radiuses of these matrices are not greater than 1. The proof given in Corrêa [10] assumes the existence of $(\nu_1, \ldots, \nu_k) \in (\mathbb{R}^m)^k$ such that $\rho[\mathscr{A}(\nu_k) \cdots \mathscr{A}(\nu_1)] > 1$ and defines an input sequence of period k such that, for initial conditions satisfying (i) or (ii), the resulting state-sequence goes unbounded.

The following lemma enables the derivation of necessary conditions for BIBS stability (from Prop. 1), which are directly stated in terms of $\{A_k: k = 0, 1, \ldots, m\}$.

LEMMA 1. *Let* $\{A_l \in \mathbb{R}^{n \times n}: l = 0, \ldots, m\}$. *If, for* $\xi \in \mathbb{R}$,

$$\forall k \in Z_+ \quad \forall(\nu_1, \ldots, \nu_k) \in (\mathbb{R}^m)^k, \quad \rho[\mathscr{A}(\nu_k) \cdots \mathscr{A}(\nu_1)] \leqq \xi,$$

then

(a) $\rho[A_0] \leqq \xi$ *and for all* $\mu \in \mathbb{R}$ *and* $l \in \{1, \ldots, m\}$, $\chi(\cdot\,; A_0 + \mu A_l) = \chi(\cdot\,; A_0)$, *where* $\chi(\cdot\,; A)$ *denotes the characteristic polynomial of* A, *and*

(b) *The semigroup generated by* $\{A_0^q A_l: l = 1, \ldots, m,\ q \in \mathbb{N}\}$ *consists* (*solely*) *of nilpotent elements.*

Proof. (a) To show that $\rho[A_0] \leqq \xi$ just take $k = 1$ and $\nu_1 = 0$. The equality of the characteristic polynomials follows from Proposition II of Corrêa [9], taking $k = 1$ and $\nu_1 = \mu e_l$, where e_l is the lth vector in the canonical basis for $\mathbb{R}^m$.

(b) The basic idea for this part of the proof is to associate to $A_0^q A_l$ the finite input sequence $u_{q,l}$ of length $q + 1$ given by $u_{q,l} = (\mu e_l, 0, \ldots, 0)$. Then to each element $A_0^{q_r} A_{l_r} \cdots A_0^{q_1} A_{l_1}$ of the semigroup generated by $\{A_0^q A_l: l = 1, \ldots, m,\ q \in \mathbb{N}\}$ we can associate the input sequence $(\nu_1, \ldots, \nu_k) = (u_{q_1,l_1}, \ldots, u_{q_r,l_r})$, thereby relating this semigroup element and the matrix $\{\mathscr{A}(\nu_k) \cdots \mathscr{A}(\nu_1)\}$. Let $r \in Z_+$, $l_1, \ldots, l_r$, $q_1, \ldots, q_r$ be given where $l_i \in \{1, \ldots, m\}$, $q_i \in \mathbb{N}$, and define $k = \sum_{i=1}^{r} (q_i + 1)$,

$$\begin{aligned} &\nu_1 = \mu e_{l_1}, \nu_2 = \cdots = \nu_{q_1+1} = 0, \nu_{q_1+2} = \mu e_{l_2}, \nu_{q_1+3} = \cdots \\ &\quad = \nu_{q_1+q_2+2} = 0, \ldots, \nu_{\sum_{i=1}^{r-1}(q_i+1)+1} = \mu e_{l_r}, \nu_{\sum_{i=1}^{r-1}(q_i+1)+2} \\ &\quad = \cdots = \nu_k = 0 \end{aligned}$$

for some $\mu \in \mathbb{R}_+$. Thus

$$\begin{aligned} \mathscr{A}(\nu_k) \cdots \mathscr{A}(\nu_1) &= A_0^{q_r}(A_0 + \mu A_{l_r}) \cdots A_0^{q_1}(A_0 + \mu A_{l_1}) \\ &= [A_0^{q_r+1} + \mu(A_0^{q_r} A_{l_r})] \cdots [A_0^{q_1+1} + \mu(A_0^{q_1} A_{l_1})] \Rightarrow \end{aligned}$$

$$\mathscr{A}(\nu_k) \cdots \mathscr{A}(\nu_1) = A_0^k + \sum_{i=1}^{r-1} M_i \mu^i + [(A_0^{q_r} A_{l_r}) \cdots (A_0^{q_1} A_{l_1})]\mu^r,$$

where $M_i \in \mathbb{R}^{n \times n}$.

Now, for all $\mu \in \mathbb{R}_+$,

$$0 \leqq \rho[\mathscr{A}(\nu_k) \cdots \mathscr{A}(\nu_1)]$$

$$= \rho\left\{\mu^r\left[(1/\mu^r)A_0^k + \sum_{i=1}^{r-1} M_i(1/\mu^{r-i}) + (A_0^{q_r}A_{l_r}) \cdots (A_0^{q_1}A_{l_1})\right]\right\} \leqq \xi \Rightarrow$$

$$0 \leqq \rho\left[(1/\mu^r)A_0^k + \sum_{i=1}^{r-1} M_i(1/\mu^{r-i}) + (A_0^{q_r}A_{l_r}) \cdots (A_0^{q_1}A_{l_1})\right] \leqq \xi/\mu^r.$$

As $\rho(\cdot)$ is continuous, taking limits as $\mu \to \infty$, it follows that

$$0 \leqq \rho[(A_0^{q_r}A_{l_r}) \cdots (A_0^{q_1}A_{l_1})] \leqq 0 \Leftrightarrow \rho[(A_0^{q_r}A_{l_r}) \cdots (A_0^{q_1}A_{l_1})] = 0. \qquad \square$$

It is perhaps worth noting that the condition on the characteristic polynomials of A_0 and $(A_0 + wA_l)$ given in (a) above is equivalent, in the case of $n = 2$, to the existence of *a* basis change, which makes *both* A_0 and A_l triangular. For $n = 3$, this equivalence does not hold, but with the addition of the requirement that $A_0^q A_l$ is nilpotent for $q = 0, 1, 2$, it also follows that A_0 and A_l can be jointly made triangular by a single basis change [9]. In general, a link between the conditions on $\{A_k: k = 0, \ldots, m\}$ given in Lemma 1 and the existence of a basis change, which makes them all triangular, is established in the following lemma.

LEMMA 2. *Let* $\{A_k: k = 0, \ldots, m\}$, $A_k \in \mathbb{R}^{n \times n}$. *If the semigroup generated by* $\{A_0^q A_l: l = 1, \ldots, m, q \in \mathbb{N}\}$ *consists of nilpotent matrices, then there exists* $T \in \mathbb{C}^{n \times n}$, *nonsingular, such that for all* $k = 1, \ldots, m$, TA_kT^{-1} *is strictly upper triangular and* TA_0T^{-1} *is upper triangular.*

Proof. An overall plan for this proof can be sketched as follows. A theorem due to Levitzky asserts the existence of a (single) similarity transformation that takes every member of the nilpotent semigroup generated by $\{A_0^q A_l: q \in \mathbb{N}, l = 1, \ldots, m\}$ to triangular form. This is then used to show that there exists a nontrivial subspace that is invariant under $A_0, A_1, \ldots, A_m$. Thus, with an appropriate basis change the problem can be reduced to an equivalent one of smaller dimension, and the proof is carried out by induction on n. Consider first the case of $n = 2$. If, for all $l = 1, \ldots, m$, $A_l = 0$, the proposition is trivially true (as any T that makes A_0 upper triangular will do). Assume, therefore, that there exists $l \in \{1, \ldots, m\}$ such that $A_l \neq 0$. For $T_2 \in \mathbb{R}^{2 \times 2}$, which takes A_l to Schur form (recall that A_l is nilpotent), let $\tilde{A}_k = T_2 A_k T_2^{-1}$, $k = 0, \ldots, m$. Then,

$$\tilde{A}_l = \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix}, \quad a \neq 0, \quad \tilde{A}_k = \begin{bmatrix} \alpha_k & \beta_k \\ \delta_k & \eta_k \end{bmatrix}, \quad \tilde{A}_k\tilde{A}_l = \begin{bmatrix} 0 & a\alpha_k \\ 0 & a\delta_k \end{bmatrix},$$

which is nilpotent. Thus $a\delta_k = 0 \Rightarrow \delta_k = 0$ (since $a \neq 0$). Hence, for all $k = 0, \ldots, m$, $\tilde{A}_k$ is upper triangular. Since, for each $k \in \{1, \ldots, m\}$, $\tilde{A}_k$ is nilpotent, it is also strictly upper triangular.

Assume now that the lemma is valid for $n \leqq n_0$ and consider $A_k \in \mathbb{R}^{(n_0+1) \times (n_0+1)}$. As the semigroup generated by $\{A_0^q A_l: l = 1, \ldots, m, q \in \mathbb{N}\}$ consists of nilpotent matrices, it follows from a theorem due to Levitzki (see Kaplansky [11, Thm. 3.5, p. 135]) that there exists $\tilde{T} \in \mathbb{R}^{(n_0+1) \times (n_0+1)}$ such that

$$\tilde{T}A_0^q A_l \tilde{T}^{-1} = \tilde{A}_0^q \tilde{A}_l \text{ is strictly upper triangular} \quad \text{for all } q \in \mathbb{N} \quad \forall l \in \{1, \ldots, m\},$$

where $\tilde{A}_k \triangleq \tilde{T}A_k\tilde{T}^{-1}$. Now write

$$\tilde{A}_0 = [\check{A}_0^T \vdots y]^T, \qquad y \in \mathbb{R}^{(n+1)}.$$

As $\tilde{A}_0(\tilde{A}_0^q\tilde{A}_l)$ is strictly upper triangular, it follows that for all $l = 1, \dots, m$,

$$y^T[\tilde{A}_l \vdots \tilde{A}_0\tilde{A}_l \vdots \cdots \vdots \tilde{A}_0^{n-1}\tilde{A}_l] = 0 \Rightarrow$$

$$y^T[\tilde{A}_{1,m} \vdots \tilde{A}_0\tilde{A}_{1,m} \vdots \cdots \vdots \tilde{A}_0^{n-1}\tilde{A}_{1,m}] = 0,$$

where

$$\tilde{A}_{1;m} \triangleq [\tilde{A}_1 \vdots \cdots \vdots \tilde{A}_m] \in \mathbb{R}^{(n_0-1)\times m(n_0+1)}.$$

If $y \neq 0$, rank $[\tilde{A}_{1,m} \vdots \cdots \vdots \tilde{A}_0^{n-1}\tilde{A}_{1,m}] = \eta < n_0 + 1$, which implies that there exists $\breve{\mathscr{U}} \in \mathbb{R}^{(n_0+1)\times\eta}$ of rank η whose columns form a basis for the column space of $[\tilde{A}_{1,m} \vdots \cdots \vdots \tilde{A}_0^{n-1}\tilde{A}_{1,m}]$. Thus, if $y \neq 0$, let $\breve{T}^{-1} = [\breve{\mathscr{U}}\ \bar{\mathscr{U}}]$, where $\breve{T}^{-1} \in \mathbb{R}^{(n+1)\times(n+1)}$ is nonsingular, and define

$$\bar{A}_k = \breve{T}\tilde{A}_k\breve{T}^{-1}(\Rightarrow \breve{T}^{-1}\bar{A}_k = \tilde{A}_k\breve{T}^{-1}) \quad \forall k = 0, \dots, m.$$

As the column space of $\breve{\mathscr{U}}$ is invariant under $\tilde{A}_0$ and contains the column space of each $\tilde{A}_l$, $\bar{A}_k$ can be written as

$$\bar{A}_0 = \left[\begin{array}{ccc} \bar{A}_{01} & \vdots & \breve{A}_0 \\ \cdots & \cdots & \cdots \\ 0 & \vdots & \bar{A}_{02} \end{array}\right] \quad \forall l = 1, \dots, m, \quad \bar{A}_l = \left[\begin{array}{ccc} \bar{A}_{l1} & \vdots & \breve{A}_l \\ \cdots & \cdots & \cdots \\ 0 & \cdots & 0 \end{array}\right], \quad \bar{A}_{k1} \in \mathbb{R}^{\eta\times\eta}.$$

If, on the other hand, $y = 0$, let $\breve{T} = I$ and $\bar{A}_k = \tilde{A}_k$. Clearly, in this case $\bar{A}_k$ also has the block-triangular structure shown above for the case of $y \neq 0$.

Now, as the semigroup generated by $\{A_0^qA_l$: $q \in \mathbb{N}$, $l = 1, \dots, m\}$ consists of nilpotent matrices, so does the semigroup generated by $\{\bar{A}_0^q\bar{A}_l$: $q \in \mathbb{N}$, $l = 1, \dots, m\}$. But as

$$\bar{A}_0^q\bar{A}_l = \left[\begin{array}{ccc} \bar{A}_{01}^q\bar{A}_{l1} & \vdots & \hat{Z}_{l,q} \\ \cdots & \cdots & \cdots \\ 0 & \vdots & 0 \end{array}\right],$$

this implies that the semigroup generated by $\{\bar{A}_{01}^q\bar{A}_{l1}$: $q \in \mathbb{N}$, $l = 1, \dots, m\}$ consists also of nilpotent matrices. Thus, it follows from the fact that $\bar{A}_{k1} \in \mathbb{R}^{\eta\times\eta}(\eta \leqq n_0)$, and from the induction hypothesis that there exists $T_\eta \in \mathbb{C}^{\eta\times\eta}$ which makes all $\{T_\eta\bar{A}_{k1}T_\eta^{-1}$: $k = 1, \dots, m\}$ strictly upper triangular and $T_\eta\bar{A}_{01}T_\eta^{-1}$ upper triangular. As a result, defining

$$\hat{T} \triangleq \left[\begin{array}{ccc} T_\eta & \vdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \vdots & \hat{T}_\eta \end{array}\right],$$

where $\hat{T}_\eta$ is nonsingular and $\hat{T}_\eta\bar{A}_{02}\hat{T}_\eta^{-1}$ is upper triangular, and $T \triangleq \hat{T}\breve{T}\tilde{T}$, it follows that for all $l = 1, \dots, m$, TA_lT^{-1} is strictly upper triangular and TA_0T^{-1} is upper triangular. □

The conditions above on $\{A_k$: $k = 0, \dots, m\}$ can be combined in such a way as to yield necessary and sufficient conditions for BIBS stability as stated in the following theorem.

THEOREM 1. *Let* $\Sigma = (\{A_k$: $k = 0, \dots, m\}, B, \bar{B})$ *denote a bilinear system as in equation* (1), *where* $\rho[A_0] \neq 1$. Σ *is* BIBS-*stable in the sense that either* (i) *or* (ii) *of Proposition* 1 *holds if and only if*

(1) $\rho[A_0] < 1$ *and for all* $l = 1, \dots, m$, $\rho[A_l] = 0$; *and*

(2) *There exists* $T \in \mathbb{C}^{n\times n}$, *nonsingular, such that for all* $k = 0, \dots, m$, TA_kT^{-1} *is upper triangular.*

Proof. The fact that (1) and (2) are necessary conditions follows from Proposition 1 and Lemma 1, and from Proposition 1 and Lemma 2, respectively. To show that (1) and (2) are sufficient for BIBS stability an additional lemma is required.

LEMMA 3. *Let $\{\mathcal{U}_k \in \mathbb{C}^{n\times n}\colon k \in Z_+\}$ be a sequence of strictly upper triangular matrices such that for all $k \in Z_+$, $\|\mathcal{U}_k\| < \Gamma \in \mathbb{R}_+$. Let $\Lambda \in \mathbb{C}^{n\times n}$ be a diagonal matrix such that $\rho(\Lambda) < 1$ and define*

$$\Pi_{k,l} \triangleq (\Lambda + \mathcal{U}_k)\cdots(\Lambda + \mathcal{U}_l), \quad k \in Z_+, \quad l \in Z_+, \quad k \geqq l.$$

Then there exists a sequence $\{\eta_r \in \mathbb{R}_+\colon r \in \mathbb{N}\}$ such that

$$\forall l \in \mathbb{N} \quad \forall k \in Z_+, \quad k > l, \quad \|\Pi_{k,l+1}\| \leqq \eta_{k-l}$$

and

$$\sum_{k=1}^{\infty} \eta_k < \infty.$$

A proof for Lemma 3 is given below. This lemma is now used to show sufficiency. To this effect write, from (2),

$$\begin{aligned} x(t) = {} & T^{-1}\{\tilde{\mathcal{A}}[w(t-1)]\cdots\tilde{\mathcal{A}}[w(0)]\}Tx(0) \\ & + T^{-1}\sum_{l=1}^{t-1}\{\tilde{\mathcal{A}}[w(t-1)]\cdots\tilde{\mathcal{A}}[w(l)]\}T[B \vdots \bar{B}]\begin{bmatrix} w(l-1) \\ u(l-1) \end{bmatrix} \\ & + [B \vdots \bar{B}]\begin{bmatrix} w(t-1) \\ u(t-1) \end{bmatrix}, \end{aligned}$$

where T is as in (2) (in the statement of Theorem 1), $\tilde{A}_k = TA_kT^{-1}$ and

$$\tilde{\mathcal{A}}(w) \triangleq \tilde{A}_0 + \sum_{l=1}^{m} w_l\tilde{A}_l.$$

Thus

$$\begin{aligned} \|x(t)\| \leqq {} & \|T^{-1}\|\left\{\xi_{t,0}\|T\|\|x(0)\| + \left(\sum_{l=1}^{t-1}\xi_{t,l}\right)\|T\|\|[B \vdots \bar{B}]\|\max\{\Gamma_w, \Gamma_u\}\right\} \\ & + \|[B \vdots \bar{B}]\|\max\{\Gamma_w, \Gamma_u\}, \end{aligned}$$

where

$$\xi_{t,l} \triangleq \|\tilde{\mathcal{A}}[w(t-1)]\cdots\tilde{\mathcal{A}}[w(l)]\| \quad \forall t \in \mathbb{N}, \quad \|w(t)\| \leqq \Gamma_w, \quad \|u(t)\| \leqq \Gamma_u.$$

Defining $\Lambda = \operatorname{diag}([\tilde{A}_0]_{11}, \ldots, [\tilde{A}_0]_{n,n})$, $\mathcal{U}_k = (\tilde{A}_0 - \Lambda) + \sum_{l=1}^{m} w_l(k-1)\tilde{A}_l$, Lemma 3 is applicable and leads to

$$\xi_{t,l} \leqq \eta_{t-l} \Rightarrow \sum_{l=1}^{t-1}\xi_{t,l} \leqq \sum_{l=1}^{\infty}\eta_l < \infty \quad \forall t \in Z_+, \quad t > 1.$$

Thus $x(\cdot)$ is bounded.

Proof of Lemma 3. The basic idea of the proof is to look at the diagonal blocks of increasing dimension of the given upper triangular product matrix $\Pi_{k,l}$ and to obtain bounds on the norm of one block, which depend solely on a bound on the norm of its preceding block and on $\rho(\Lambda)$. Since the norm of the top 1×1 block is bounded by

$\rho(\Lambda)^{k-l}$, this iteration leads to the desired summable bounds on $\|\Pi_{k,l}\|$. For $n = 1$, $\mathcal{U}_k = 0$ and it suffices to make $\eta_{k-l} = \Lambda^{k-l}$. Suppose then that $n > 1$ and, for $q = 1, \ldots, n$, define

$$\Lambda_q \triangleq [I_q \vdots 0]\Lambda\begin{bmatrix} I_q \\ \cdots \\ 0 \end{bmatrix}, \qquad \mathcal{U}_k^q = [I_q \vdots 0]\mathcal{U}_k\begin{bmatrix} I_q \\ \cdots \\ 0 \end{bmatrix},$$

where I_q is the $q \times q$ identity matrix and

$$\Pi_{k,l+1}^q = (\Lambda_q + \mathcal{U}_k^q) \cdots (\Lambda_q + \mathcal{U}_{l+1}^q).$$

Note that $\Pi_{k,l+1}^n = \Pi_{k,l+1}$,

$$\mathcal{U}_k^{q+1} = \begin{bmatrix} \mathcal{U}_k^q & \vdots & u_k^q \\ \cdots & \cdots & \cdots \\ 0 & \vdots & 0 \end{bmatrix}, \quad \Lambda_{q+1} = \begin{bmatrix} \Lambda_q & \vdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \vdots & \lambda_{q+1} \end{bmatrix},$$

$$\Pi_{k,l+1}^{q+1} = \begin{bmatrix} \Pi_{k,l+1}^q & \vdots & y_{k,l+1}^q \\ \cdots & \cdots & \cdots \\ 0 & \vdots & \lambda_{q+1}^{k-l} \end{bmatrix},$$

where

$$y_{k,l+1}^q = \sum_{i=1}^{(k-l)-1} \Pi_{k,l+i+1}^q u_{l+i}^q \lambda_{q+1}^{i-1} + \lambda_{q+1}^{(k-l)-1} u_k^q.$$

It then follows that

$$\|\Pi_{k,l+1}^{q+1}\| \leqq \|\Pi_{k,l+1}^q\| + \|y_{k,l+1}^q\| + |\lambda_{q+1}|^{k-l} \Rightarrow$$

$$\|\Pi_{k,l+1}^{q+1}\| \leqq \|\Pi_{k,l+1}^q\| + \rho(\Lambda)^{k-l} + \Gamma\rho(\Lambda)^{k-l-1} + \Gamma \sum_{i=1}^{k-l-1} \rho(\Lambda)^{i-1}\|\Pi_{k,l+i+1}^q\|.$$

Assume, for a moment, that there exists a sequence $\{\eta_r^q \in \mathbb{R}_+ : r \in \mathbb{N}\}$ such that for all $l \in \mathbb{N}$ and $k \in Z_+$, $k > l$, $\|\Pi_{k,l+1}^q\| \leqq \eta_{k-l}^q$ and $\sum_{k=1}^{\infty} \eta_k^q < \infty$. Then,

$$\|\Pi_{k,l+1}^{q+1}\| \leqq \eta_{k-l}^q + [\rho(\Lambda)^{k-l} + \Gamma\rho(\Lambda)^{k-l-1}] + \Gamma\alpha_{k-l}^q,$$

where

$$\alpha_t^q \triangleq \sum_{i=1}^{t-1} \rho(\Lambda)^{i-1}\eta_{t-i}^q,$$

so that defining

$$\eta_t^{q+1} = \eta_t^q + [\rho(\Lambda)^t + \Gamma\rho(\Lambda)^{t-1}] + \Gamma\alpha_t^q, \quad t \in Z_+,$$

it follows that $\|\Pi_{k,l+1}^{q+1}\| \leqq \eta_{k-l}^{q+1}$ and $\sum_{t=1}^{\infty} \eta_t^{q+1} < \infty$, since $\sum_{t=1}^{\infty} \alpha_t^q < \infty$ (this, in turn, is due to $\sum_{k=1}^{\infty} \eta_k^q < \infty$, $\rho(\Lambda) < 1$; compare Dunford and Schwartz [12, p. 529]). As the existence of $\{\eta_k^1 : k \in \mathbb{N}\}$ has been asserted above (i.e., $\eta_k^1 = \rho(\Lambda)^k$), this argument can be repeatedly applied for $q = 1, \ldots, n-1$, which concludes the proof. □

Remark. Lemma 3 remains valid if Λ is replaced by a sequence of diagonal matrices Λ_k such that $\rho(\Lambda_k) < \lambda < 1$ and $\Pi_{k,l}$ is defined as

$$\Pi_{k,l} = (\Lambda_k + \mathcal{U}_k) \cdots (\Lambda_l + \mathcal{U}_l).$$

The proof, in this case, follows exactly the same argument as presented above. The simpler version was proven here to avoid more involved notation.

For a given bilinear system $\Sigma = (\{A_k : k = 0, \ldots, m\}, B, \bar{B})$, which is BIBS-stable in the sense of (i) or (ii) of Proposition 1, Theorem 1 says nothing about which specific basis change takes all $\{A_k : k = 0, \ldots, m\}$ simultaneously to triangular form. In fact, it is possible to be more specific about such basis changes by introducing an assumption on the matrices $\{A_l : l = 1, \ldots, m\}$, as done in the following theorem.

THEOREM 2. *Let* $\Sigma = (\{A_k : k = 0, \ldots, m\}, B, \bar{B})$ *denote a bilinear system as in equation* (1), *where* $\rho(A_0) \neq 1$. *Assume that there exists* $i \in \{1, \ldots, m\}$ *such that* rank $(A_i) \geqq n - 1$. *Then,* Σ *is* BIBS-*stable in the sense that either* (i) *or* (ii) *holds if and only if*

(1) $\rho[A_0] < 1$ *and for all* $l = 1, \ldots, m$, $\rho[A_l] = 0$;

(2′) *If* $T \in \mathbb{C}^{n \times n}$ *is a nonsingular matrix such that* TA_iT^{-1} *is in* (*upper triangular*) *Jordan form, then* TA_kT^{-1} *is upper triangular for all* $k = 0, 1, \ldots, m$.

Proof. It suffices to show that (2′) is a necessary condition for BIBS stability. To this effect recall that, as a consequence of Proposition 1 and Lemma 1, the semigroup generated by $\{A_0^q A_l : l = 1, \ldots, m, q \in \mathbb{N}\}$ consists of nilpotent matrices. Hence, for all $k = 0, \ldots, m$ and $r \in Z_+$, $A_k A_i^r$ is nilpotent. Now let $\tilde{A}_k = TA_kT^{-1}$, $k = 0, 1, \ldots, m$. As $\rho[A_i] = 0$, rank $(A_i) = n - 1$ and $\tilde{A}_i$ is in upper triangular Jordan form,

$$\tilde{A}_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \\ \vdots & & & & \\ \vdots & & & & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}.$$

Let $\tilde{A}_k = (\tilde{a}_{rs})$. Notice now that

$$\tilde{A}_i^{n-1}\tilde{A}_k = \begin{bmatrix} \tilde{a}_{n,1} & \cdots & \tilde{a}_{n,n} \\ 0 & & 0 \\ \vdots & & \vdots \\ 0 & & 0 \end{bmatrix}, \quad \tilde{A}_i^{n-2}\tilde{A}_k = \begin{bmatrix} \tilde{a}_{n-1,1} & \cdots & \tilde{a}_{n-1,n} \\ \tilde{a}_{n,1} & \cdots & \tilde{a}_{n,n} \\ 0 & & 0 \\ \vdots & & \vdots \\ 0 & & 0 \end{bmatrix},$$

$$\tilde{A}_i^{n-l}\tilde{A}_k = \begin{bmatrix} \tilde{a}_{n-l+1,1} & \cdots & \tilde{a}_{n-l+1,n} \\ \cdots & & \cdots \\ \tilde{a}_{n,1} & \cdots & \tilde{a}_{n,n} \\ 0 & \cdots & 0 \end{bmatrix} \quad \forall l = 1, \ldots, n-1.$$

Clearly, the conditions $\tilde{A}_i^{n-1}\tilde{A}_k$ nilpotent, $\tilde{A}_i^{n-2}\tilde{A}_k$ nilpotent, $\ldots$, $\tilde{A}_i\tilde{A}_k$ nilpotent imply, respectively, that $\tilde{a}_{n,1} = 0$; $\tilde{a}_{n-1,1}$ and $\tilde{a}_{n,2} = 0$; $\ldots$; $\tilde{a}_{n-l+1,1} = 0, \ldots, \tilde{a}_{n,l} = 0$; $\ldots$; $\tilde{a}_{2,1} = 0, \ldots, \tilde{a}_{n,n-1} = 0$. Thus, $\tilde{A}_k$ is upper triangular. □

Finally, it may be worthwhile to comment on the asymptotic stability of the state affine systems defined by equation (1) for each pair of bounded sequences (w, u), which is denoted by $\Sigma(w, u)$. To this effect consider the standard definition below.

DEFINITION 2. $\Sigma(w, u)$ is said to be asymptotically stable if and only if every solution $x(\cdot, x_0)$ given by equation (2) is asymptotically stable, i.e., for all $x_0 \in \mathbb{R}^n$, there exists a neighborhood of x_0, say $V(x_0)$, such that for all $\tilde{x} \in V(x_0)$, $\|x(t, \tilde{x}) - x(t, x_0)\| \to 0$ as $t \to \infty$.

The following theorem is a simple consequence of the lemmas given above.

THEOREM 3. *Let* $\{A_k : k = 0, \ldots, m\}$, $A_k \in \mathbb{R}^{n \times n}$ *be given. The following propositions are equivalent.*

(I) *For any pair of bounded sequences* (w, u), $\Sigma(w, u)$ (*as given by equation* (1)) *is asymptotically stable.*

(II) *Conditions* (1) *and* (2) *of Theorem* 1 *hold.*

Proof. Note first that (I) holds if and only if, for all $\tilde{x}$ in a neighborhood of 0, $\|\mathscr{A}_t(w)\tilde{x}\| \rightarrow 0$ as $t \rightarrow \infty$, where

$$\mathscr{A}_t(w) \triangleq \mathscr{A}[w(t-1)] \cdots \mathscr{A}[w(0)].$$

It is now shown that (I) implies (II). To this effect, note first that using essentially the same argument as in the proof of Proposition 1, it can be shown that

$$(\mathrm{I}) \Rightarrow \forall k \in Z_+, \quad \forall (\nu_1, \ldots, \nu_k) \in \mathbb{R}^{km}, \quad \rho[\mathscr{A}(\nu_k) \cdots \mathscr{A}(\nu_1)] < 1,$$

from which it follows that $\rho[A_0] < 1$, and that (through Lemma 1) for all $l = 1, \ldots, m$, $\rho[A_l] = 0$. And, using Lemmas 1 and 2, it follows that (I) implies that condition (2) of Theorem 1 also holds.

It is now shown that (II) implies (I). Proceeding along the same lines as in the proof of Theorem 1, Lemma 3 can be used to show that $\|\mathscr{A}_k(w)\| \leqq \eta_k$, where $\sum_{k=1}^{\infty} \eta_k < \infty$. Thus $\eta_k \rightarrow 0$, from which it follows that

$$\forall \tilde{x} \in \mathbb{R}^n, \quad \|\mathscr{A}_t(w)\tilde{x}\| \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty. \qquad \square$$

3. Concluding remarks. Bounded-input/bounded-state stability of discrete-time, multi-input bilinear systems was addressed in this article. First, a necessary condition for a system $\Sigma = (\{A_k: k = 0, \ldots, m\}, B)$ to be BIBS-stable was stated in terms of the spectral radiuses of all matrices in the range of a function from $\mathbb{R}^k$ to $\mathbb{R}^{n \times n}$ being smaller or equal to one, for any positive integer k. Then two lemmas established the equivalence of this condition and the *joint* triangularizability (via similarity transformation) of $\{A_k: k = 0, \ldots, m\}$ together with the following constraints on their spectral radiuses:

$$\rho(A_0) \leqq 1, \quad \rho(A_k) = 0, \quad k = 1, \ldots, m.$$

On the basis of another lemma it was then shown that, whenever $\rho[A_0] \neq 1$, these conditions are indeed sufficient.

REFERENCES

[1] C. Bruni, G. Di Pillo, and G. Koch, *Bilinear systems: An appealing class of "nearly linear" systems in theory and applications*, IEEE Trans. Automat. Control, Vol AC-19 (1974), pp. 334–348.

[2] R. R. Mohler and W. J. Kolodziej, *An overview of bilinear system theory and applications*, IEEE Trans. Systems Man. Cybernet, SMC-10 (1980), pp. 633–688.

[3] T. Goka, T. J. Tarn, and J. Zabersky, *On the controllability of a class of discrete bilinear systems*, Automatica, 9 (1973), pp. 615–622.

[4] Stensholt and D. Tjostheim, *Multiple bilinear time-series models*, J. Time Ser. Anal., 8 (1987), pp. 221–233.

[5] T. Subba Rao, *Spectral and bispectral methods for the analysis of nonlinear time-series signals*, in Nonlinear Time Series and Signal Processing, R. R. Mohler, ed., Lecture Notes in Control 106, Springer-Verlag, Berlin, New York, 1987.

[6] F. Fnaiech and L. Ljung, *Recursive identification of bilinear systems*, Internat J. Control, 45 (1987), pp. 453–470.

[7] C. S. Kubrusly and O. L. V. Costa, *Mean square stability conditions for discrete-time, stochastic bilinear systems*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1082–1087.

[8] C. Gounaridis-Minadis and N. Kalouptsidis, *Stability of discrete-time bilinear systems with constant input*, Internat. J. Control, 43 (1986), pp. 663–669.

[9] G. Corrêa, *On bounded-input/bounded-state stability of discrete-time, bilinear systems*, Res. Rep. 013/90, LNCC/CNPq, Rio de Janeiro, 1990.

[10] G. Corrêa, *Necessary and sufficient conditions for bounded-input/bounded-output stability of discrete-time, bilinear systems*, Res. Rep. 015/91, LNCC/CNPq, Rio de Janeiro, 1991.

[11] I. Kaplansky, *Fields and Rings*, The University of Chicago Press, Chicago, IL, 1972.

[12] N. Dunford and J. T. Schwartz, *Linear Operators-Part 1: General Theory*, Wiley-Interscience, New York, 1958.

[13] W. Rudin, *Principles of Mathematical Analysis*, Mc-Graw Hill, New York, 1973.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 721-734, July 1993

010

APPROXIMATION BY A HERMITIAN POSITIVE SEMIDEFINITE TOEPLITZ MATRIX*

T. J. SUFFRIDGE† AND T. L. HAYDEN†

Abstract. The problem of finding the nearest Hermitian positive semidefinite Toeplitz matrix (in the Frobenius norm) of a given rank to an arbitrary matrix is considered. A special orthogonal basis and equally spaced frequencies allow good initial approximations. A second method using alternating projections solves the case of unrestricted rank. Some interesting numerical results suggest possible applications to signal processing problems.

Key words. Toeplitz matrices, positive definite Hermitian Toeplitz approximation, self-inversive polynomials

AMS subject classifications. 15A48, 15A57, 30EXX

1. Introduction. This paper addresses the problem of best approximation of a given matrix by a positive semidefinite Hermitian Toeplitz matrix. Related problems occur in many engineering and statistical applications [3], especially in the area of signal processing. Cybenko [4] considered problems in this area and related his results to moment problems. For example, he showed how one may embed an $n \times n$ Toeplitz matrix of rank r into the family of $n + 1 \times n + 1$ Toeplitz matrices of rank r using the Levinson–Durbin algorithm. Other Toeplitz approximation problems are discussed by Cybenko [4] and Kung [13].

We consider the following problem. Let $\mathscr{T}_n^r$ denote the set of $n \times n$ positive semidefinite Hermitian Toeplitz matrices of rank $\leqq r$. Let D be a given complex $n \times n$ (data) matrix.

PROBLEM 1.

$$\min_{T \in \mathscr{T}_n^r} \|D - T\|.$$

Here $\|\cdot\|$ is the Frobenius matrix norm given by $\|A\|^2 = \langle A, A \rangle$ and the inner product $\langle A, B \rangle = \operatorname{trace}(B^*A) = \sum_{i,i} a_{ij}\bar{b}_{ij}$, where the bar denotes complex conjugation, and $B^* = \bar{B}^T$.

This problem and other approximation problems were studied by Shaw and Kumaresan [15] in relation to signal processing problems. We do not attempt to directly relate our results to signal processing; however, we will make a partial connection through the following factorization.

If T is a Hermitian positive semidefinite Toeplitz matrix, then there exists a factorization

$$T = VDV^*, \tag{1}$$

where V is an $n \times q$ Vandermonde matrix ($q \leqq n$) and D is a diagonal matrix with positive numbers on the diagonal. This decomposition follows from Caratheodory's theorem on trigonometric moment problems [10, Thm. 12, p. 24]. Since the factorization is a key component of our results, we wish to make the connection clear with Caratheodory's theorem, which is proved by the classical techniques of moments.

* Received by the editors February 13, 1991; accepted for publication (in revised form) October 21, 1991. This work was supported in part by National Science Foundation grant CHE-9005960.

† Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506 (ted@ms.uky.edu and hayden@ms.uky.edu).

THEOREM (CARATHEODORY). *If $c_0, c_1, \ldots, c_{n-1}$ is a nonnegative definite sequence with rank $q < n$, then there exists a unique canonical representation*

$$c_k = \sum_{j=1}^{q} \rho_j \alpha_j^k, \qquad k = 0, \pm 1, \ldots, \pm(n-1)$$

in which each $\rho_j > 0$, $|\alpha_j| = 1$ for all j, $1 \leqq j \leqq q$, and the α_j are distinct. In this representation, q = rank of the Toeplitz form $\sum_{k,j=0}^{n-1} c_{k-j} z_k \bar{z}_j$.

To obtain the factorization (1) we make the following observation. The statement that $c_0, c_1, \ldots, c_{n-1}$ is a nonnegative definite sequence of rank q means that the Hermitian Toeplitz matrix $T = (c_{j-k})$, $c_{-l} = \bar{c}_l$, $0 \leqq l \leqq n-1$ is positive semidefinite of rank q. Further, if D is the $q \times q$ diagonal matrix with diagonal $\rho_j > 0$, $1 \leqq j \leqq q$, and the jth column of V is $[1, \bar{\alpha}_j, \bar{\alpha}_j^2, \ldots, \bar{\alpha}_j^{n-1}]^T$, where $|\alpha_j| = 1$, $1 \leqq j \leqq q$, then $T = (c_{j-k})$ where $c_k = \sum_{j=1}^{q} \rho_j \alpha_j^k$. Hence T has the factorization (1) with positive entries on the $q \times q$ diagonal matrix D and distinct $e^{i\theta_k}$ forming the columns of the Vandermonde matrix V if and only if T is a Hermitian positive semidefinite Toeplitz matrix. The representation is unique if $q < n$.

The columns of V are of the form $[1, e^{i\theta_k}, e^{2i\theta_k}, \ldots, e^{(n-1)i\theta_k}]^T$, $k = 1, 2, \ldots, q$, and each θ_k can represent a frequency one may associate with a signal processing problem. The θ values in V turn out to yield the roots of Szegö polynomials. The theory for a computational solution of (1) was developed by Gragg [9]. Basically one first computes the Schur parameters (reflection coefficients, partial correlation coefficients). From the Schur parameters a certain Hessenberg matrix is formed. The eigenvalues of this matrix give the frequencies θ_k, while the values of the matrix D are obtained from the first component of their eigenvectors. The basis of these steps lies in the relationship of the Szegö polynomials and an analogue of the Christoffel–Darboux formula to the development of a transliteration of the Levinson algorithm from the space of polynomials to an isometric operator on a Hilbert space, yielding an isometric analogue of the Hermitian Lanczos process. Based on this theory, a number of papers by Ammar, Gragg, and Reichel have developed fast algorithms for solving facets of this theory associated with signal processing. See, for example, [1] and [14] and references therein.

In this paper we introduce two new methods to solve Problem 1. For low rank approximations, we first develop a useful orthogonal basis for the space $\mathscr{T}_n$ of $n \times n$ Hermitian Toeplitz matrices. Using this basis, we find that the inner product of a member of $\mathscr{T}_n$ and an element of the basis can be expressed in terms of a self-inversive polynomial $P(z)$. Finding the roots of the derivative of $P(z)/z^{n-1}$ enables one to approximate the dominant frequency in the data matrix. The basic theory and orthogonal basis are developed in § 2. The rank-1 approximation occurs in § 3, while extensions of the ideas developed for the rank-1 case are given in § 4 for higher rank cases. Section 5 contains numerical results using this approach.

The second method uses the ideas of a modified alternating projection algorithm which was successful in solving similar approximation problems for distance matrices [7]. This method only solves the case of unrestricted rank, that is, the case where the minimization in Problem 1 is over the set $\mathscr{T}_n^{n-1}$. The method depends on expressing the convex set $\mathscr{T}_n^n$ as the intersection of the convex set of the cone of positive semidefinite matrices and the subspace of Hermitian Toeplitz matrices. Since $\mathscr{T}_n^{n-1}$ is the boundary of $\mathscr{T}_n^n$, Problem 1 has a unique solution when D is not positive semidefinite, which the modified alternating projection algorithm computes. For $r < n-1$, $\mathscr{T}_n^r$ is only a subset of the boundary of the convex set $\mathscr{T}_n^n$, and this drastically increases the difficulty of the lower rank problems. There is a unique minimum for $r = n-1$, but many local maxima, minima, and saddle points when $r < n - 1$. However, approximations in $r = n - 1$ can sometimes be useful for lower rank approximations. This approach has been fruitful in the distance matrix approximation problem (see, for example, [8]), and we plan to pursue the idea in a future work.

2. Basic results. The existence and uniqueness of solutions to Problem 1 are given in Theorem 1.

THEOREM 1. *Problem* 1 *has a unique solution for* $r = n$ *and has a unique solution for* $r = n - 1$ *provided the data matrix is not positive semidefinite. In all other cases there exists a solution, which may not be unique.*

Proof. Since $\mathcal{T}_n^n$ is the intersection of the closed convex set of positive semidefinite matrices and the subspace of Hermitian Toeplitz matrices, $\mathcal{T}_n^n$ is a closed convex cone. If $D \in \mathcal{T}_n^n$, then the solution is D. Otherwise the solution is the near point in $\mathcal{T}_n^n$ to D, which is unique by convexity. Since $\mathcal{T}_n^{n-1}$ is the boundary of $\mathcal{T}_n^n$, if D is not positive semidefinite, the unique near point to D in $\mathcal{T}_n^n$ will lie in $\mathcal{T}_n^{n-1}$ and provide the unique solution. In all other cases $\mathcal{T}_n^r$ is closed, but is not convex if $r < n - 1$. Intersect $\mathcal{T}_n^r$ with a closed sphere S with center 0 and radius $\|D\|$. Then $\mathcal{T}_n^r \cap S$ is a closed compact set, and the continuous function $\|D - T\|$ clearly has a minimum on this set.

After we develop a basis for the Hermitian Toeplitz matrices, it is easy to construct examples where the solution is not unique for $r < n - 1$. See Example 1 in § 3.

In order to solve Problem 1, we establish a preliminary result that allows the arbitrary $n \times n$ data matrix D to be replaced by a certain Hermitian Toeplitz matrix A with the same resulting solution. Hence it will suffice to only consider data matrices which are Hermitian and Toeplitz. Let $\mathcal{H}_n$ denote the family of $n \times n$ Hermitian matrices, and let A^* denote the complex conjugate of A^T, so $A \in \mathcal{H}_n$ implies $A = A^*$. In Lemma 1, $\|\cdot\|$ is the Frobenius norm and $\langle , \rangle$ is the inner product generating the norm as given in the introduction.

LEMMA 1. *Let* A *be a* $n \times n$ *matrix and let* $H \in \mathcal{H}_n$, $B = (A + A^*)/2 \in \mathcal{H}_n$. *Then*

$$\|A - H\|^2 = \|A - B\|^2 + \|B - H\|^2.$$

Proof. The proof follows from the fact that $\operatorname{Re}\langle A - B, B - H\rangle = 0$.

Now let $A = [a_{ij}]$ be an arbitrary $n \times n$ matrix and $T = [t_{ij}]$ be an $n \times n$ Toeplitz matrix. Then

$$\langle A, T\rangle = t_{11} \sum_{j=1}^{n} a_{j,j} + \sum_{l=1}^{n-1} \bar{t}_{1,1+l} \sum_{j=1}^{n-l} a_{j,j+l} + \sum_{l=1}^{n-1} \bar{t}_{1+l,1} \sum_{j=1}^{n-l} a_{j+l,j}.$$

Now let $\hat{A} = (\alpha_{jk})$ be the average of the diagonals of A, that is,

$$(2) \qquad \alpha_{q,q+j} = \frac{1}{n-j} \sum_{l=1}^{n-j} a_{l,j+1} \quad \text{and} \quad \alpha_{q+j,q} = \frac{1}{n-j} \sum_{l=1}^{n-j} a_{j+l,l}$$

for $0 \leqq j \leqq n - q$. Then it is clear that

(a) $\hat{A}$ is Toeplitz;

(b) $\langle A, T\rangle = \langle \hat{A}, T\rangle$ for all Toeplitz matrices T.

Furthermore, for $S \in \mathcal{T}_n^p$,

$$\text{(c)} \quad \|A - S\|^2 = \|A - \hat{A}\|^2 + 2 \operatorname{Re}\langle A - \hat{A}, \hat{A} - S\rangle + \|\hat{A} - S\|^2$$
$$= \|A - \hat{A}\|^2 + \|\hat{A} - S\|^2$$

because $\hat{A} - S$ is Toeplitz. These observations prove the following lemma.

LEMMA 2. *Let $A = [a_{jk}]$ be a given $n \times n$ matrix and define $\hat{A}$ by* (2). *For all Toeplitz matrices T, $\langle A, T\rangle = \langle \hat{A}, T\rangle$ and $\|A - T\|^2 = \|A - \hat{A}\|^2 + \|\hat{A} - T\|^2$. Hence* $\min\{\|A - T\|$: *T is Toeplitz*$\} = \|A - \hat{A}\|$.

Lemmas 1 and 2 combine to show that one can reduce Problem 1 to the case of Hermitian Toeplitz data matrices; see also [5].

THEOREM 2. *Let $A = [a_{ij}]$ be an arbitrary $n \times n$ matrix, and let $\hat{A}$ be given by* (2). *Then $T \in \mathscr{T}_n^p$ satisfies*

$$\|A - T\| = \min_{S \in T_n^p} \|A - S\|$$

if and only if

$$\|(\hat{A} + \hat{A}^*)/2 - T\| = \min_{S \in T_n^p} \|(\hat{A} + \hat{A}^*)/2 - S\|.$$

We next find a basis for the $n \times n$ Hermitian Toeplitz matrices $\mathscr{T}_n$. Now the members of $\mathscr{T}_n$ have complex entries, but we view $\mathscr{T}_n$ as a vector space over the reals in order that $\mathscr{T}_n$ be closed under scalar multiplication. Hence we write $\mathscr{T}_n = \mathscr{R}_n + i\mathscr{S}_n$ where each $R \in \mathscr{R}_n$ is a real symmetric Toeplitz matrix and each $S \in \mathscr{S}_n$ is a real skew-symmetric Toeplitz matrix. Hence each $A \in \mathscr{T}_n$ can be written as $A = R + iS$ for $R \in \mathscr{R}_n$ and $S \in \mathscr{S}_n$. Then $\dim(\mathscr{R}_n) = n$, $\dim(S_n) = n - 1$, and $\dim(\mathscr{T}_n) = 2n - 1$. The following lemma is helpful in finding an orthogonal basis for $\mathscr{T}_n$.

LEMMA 3. *Let $T = [t_{jk}]$ and $S = [s_{jk}]$ be $n \times n$ matrices with $t_{jk} = te^{i(j-k)\theta}$ and $s_{jk} = se^{i(j-k)\phi}$. Then*

$$\langle T, S\rangle = \text{st}\,\frac{1 - \cos n(\theta - \phi)}{1 - \cos(\theta - \phi)} \qquad (= n^2 \text{ st if } \theta = \phi).$$

Proof. Let $\alpha = \theta - \phi$ and note that $\langle T, S\rangle = \text{st} \sum_{j,k=1}^{n} e^{i(j-k)\alpha}$. The proof is by induction. The result is clear for $n = 1$, so we assume it is true for $n - 1$. Then

$$\sum_{j,k=1}^{n} e^{i(j-k)\alpha} = \frac{1 - \cos(n-1)\alpha}{1 - \cos\alpha} + \sum_{j=1}^{n} e^{i(j-n)\alpha} + \sum_{k=1}^{n-1} e^{i(n-k)\alpha}$$
$$= \frac{1 - \cos(n-1)\alpha}{1 - \cos\alpha} + \sum_{j=1-n}^{n-1} e^{ij\alpha}.$$

The last sum is

$$\frac{e^{-i(n-1)\alpha} - e^{in\alpha}}{1 - e^{i\alpha}} = \frac{(e^{-i(n-1)\alpha} - e^{in\alpha})(1 - e^{-i\alpha})}{(1 - e^{i\alpha})(1 - e^{-i\alpha})}$$
$$= \frac{2\cos(n-1)\alpha - 2\cos n\alpha}{2 - 2\cos\alpha}.$$

Hence the lemma follows. □

Now set $\gamma = e^{2\pi i/n}$ so that $\gamma^n = 1$ and, furthermore, $\gamma^l = 1$ if and only if l is an integer multiple of n.

LEMMA 4. *For $0 \leqq l \leqq n - 1$, set*

$$T_l = \frac{1}{n}[\gamma^{l(j-k)}]_{j,k=1}^{n}. \tag{3}$$

Then $\{T_l: 0 \leqq l \leqq n - 1\}$ is an orthonormal subset of $\mathscr{T}_n$, and $T_lT_q = 0$ for $l \neq q$ and $T_lT_l = T_l$.

Proof. The orthogonality follows from Lemma 3. The other assertions follow from direct computations using the fact that

$$\sum_{j=1}^{n} \gamma^{lj} = \begin{cases} 0 & \text{if } l \text{ is not a multiple of } n, \\ n & \text{if } l \text{ is a multiple of } n. \end{cases} \qquad \square$$

Now define R_l, $1 \leqq l < n/2$, and S_l, $1 \leqq l \leqq n/2$, as follows:

$$R_l = \frac{1}{c}\left[\begin{pmatrix} 0 & -lI_{n-l} \\ (n-l)I_l & 0 \end{pmatrix} + \begin{pmatrix} 0 & (n-l)I_l \\ -lI_{n-l} & 0 \end{pmatrix}\right],$$

$$(4) \qquad \begin{aligned} S_l &= \frac{1}{c}\left[\begin{pmatrix} 0 & -ilI_{n-l} \\ i(n-l)I_l & 0 \end{pmatrix} + \begin{pmatrix} 0 & -i(n-l)I_l \\ ilI_{n-l} & 0 \end{pmatrix}\right], \qquad l \neq \frac{n}{2}, \\ &= \frac{1}{\sqrt{n}}\begin{pmatrix} 0 & -iI_l \\ iI_l & 0 \end{pmatrix} \quad \text{if } l = n/2, \end{aligned}$$

where $c = [2nl(n-l)]^{1/2}$.

It is easy to see that $\langle R_l, R_q \rangle = 0$ and $\langle S_l, S_q \rangle = 0$ if $l \neq q$. Further, since each of R_l and S_q is Hermitian, $\langle R_l, S_q \rangle$ is real. But since the entries of R_l are real and those of S_q are pure imaginary, $\langle R_l, S_q \rangle$ is pure imaginary. Thus $\langle R_l, S_q \rangle$ must be zero. The factor c is chosen to make $\|R_l\| = \|S_l\| = 1$. Now observe that

$$nc\langle T_l, R_q \rangle = q(n-q)[-\gamma^{-lq} + \gamma^{-l(n-q)} - \gamma^{lq} + \gamma^{l(n-q)}].$$

Since $\gamma^{ln} = \gamma^{-ln} = 1$, we see that $\langle T_l, R_q \rangle = 0$ and similarly $\langle T_l, S_q \rangle = 0$ for each l and q. This gives the basis for $\mathcal{T}_n$.

THEOREM 3. *If* $\{T_l\}$, $\{R_l\}$, *and* $\{S_l\}$ *are given by* (3) *and* (4), *then* $\{T_l: 0 \leqq l \leqq n-1\} \cup \{R_l: 1 \leqq l < n/2\} \cup \{S_l: 1 \leqq l \leqq n/2\}$ *is an orthonormal basis for* $\mathcal{T}_n$.

Note that if $A = [a_{jk}]$ and $B = [b_{jk}]$ and for a fixed θ,

$$(5) \qquad A(\theta) = [a_{jk}e^{i(j-k)\theta}] \quad \text{and} \quad B(\theta) = [b_{jk}e^{i(j-k)\theta}],$$

then $\langle A, B \rangle = \langle A(\theta), B(\theta) \rangle$. This observation gives the following corollary.

COROLLARY 1. *For fixed* θ,

$$\{T_l(\theta): 0 \leqq l \leqq n-1\} \cup \{R_l(\theta): 1 \leqq l < n/2\} \cup \{S_l(\theta): 1 \leqq l \leqq n/2\}$$

is an orthonormal basis for $\mathcal{T}_n$.

3. Rank-1 case. We now consider Problem 1 for $r = 1$, that is,

$$\min_{T \in T^1_n} \|D - T\|.$$

By Theorem 2 we may assume $D \in \mathcal{T}_n$, and we see $T \in \mathcal{T}^1_n$ must be of the form $T = t[e^{i(j-k)\theta}]$ for some θ and $t \geqq 0$, that is, $T = ntT_0(\theta)$, where $T_0(\theta)$ is given by (3). Using the fact that $\mathcal{T}_n$ is an inner product space and the fact that $\|T_0(\theta)\| = 1$, we know that for fixed θ, $\|D - T\|$ will be smallest when $nt = \langle D, T_0(\theta) \rangle$ (which is real) and in that case $\|D - T\|^2 = \|D\|^2 - |\langle D, T_0(\theta) \rangle|^2 = \|D\|^2 - n^2t^2$. Thus the problem is to find $\max_\theta \langle D, T_0(\theta) \rangle$. If D is merely Hermitian and Toeplitz (not positive semidefinite), then the above maximum may be negative. In that case, the O matrix is nearer to D than every $T \in \mathcal{T}^1_n$. However, we show in Theorem 4 below that if D is positive semidefinite, then $\langle D, T_0(\theta)\rangle \geqq 0$ for ever θ. For motivation we first consider the case $n = 2$. So

$$D = \begin{pmatrix} u & ve^{-i\phi} \\ ve^{i\phi} & u \end{pmatrix}$$

where u is real and $v \geqq 0$. Then $\langle A, T_0(\theta)\rangle = v\cos(\theta - \phi) + u$, which is greatest when $\theta = \phi$. From the basic properties of inner product spaces the following results are easily obtained.

(i) If $u + v \leqq 0$, then $\|D\| \leqq \|D - T\|$ for all $T \in \mathscr{T}_2^1$, so 0 is the best approximation to A in $\mathscr{T}_2^1$.

(ii) If $u + v > 0$ then

$$T = \frac{u+v}{2}\begin{pmatrix} 1 & e^{-i\phi} \\ e^{i\phi} & 1 \end{pmatrix} \quad \text{and} \quad D - T = \frac{u-v}{2}\begin{pmatrix} 1 & -e^{-i\phi} \\ e^{i\phi} & 1 \end{pmatrix}.$$

In this case, if rank $D = 1$ (so $u = v > 0$), then $D = T$.

(iii) If D is positive semidefinite ($u \geq v$), then

$$\begin{aligned} D &= \frac{u+v}{2}\begin{pmatrix} 1 & e^{-i\phi} \\ e^{i\phi} & 1 \end{pmatrix} + \frac{u-v}{2}\begin{pmatrix} 1 & -e^{-i\phi} \\ -e^{i\phi} & 1 \end{pmatrix} \\ &= (u+v)T_0(\phi) + (u-v)T_1(\phi), \end{aligned}$$

where T_0 and T_1 are given by (3) with $n = 2$.

In general, to find the best rank-1 approximation to $A \in \mathscr{T}_n$, $A = [\alpha(j-k)]$, $\alpha(j-k) = \overline{\alpha(k-j)}$, we require θ so that $\langle A, T_0(\theta)\rangle$ is a maximum. Now $\langle A, T_0(\theta)\rangle$ is a trigonometric polynomial of degree $n - 1$ in θ, which can be expressed in terms of the following self-inversive polynomial. Let

$$\begin{aligned} P(z) &= \overline{\alpha(n-1)}z^{2n-2} + 2\overline{\alpha(n-2)}z^{2n-3} + \cdots + (n-1)\overline{\alpha(1)}z^n \\ &\quad + n\alpha(0)z^{n-1} + (n-1)\alpha(1)z^{n-2} + \cdots + 2\alpha(n-2)z + \alpha(n-1). \end{aligned} \tag{6}$$

Then $P(z) = z^{2(n-1)}\bar{P}(1/z)$ (i.e., P is self-inversive) and $\langle A, T_0(\phi)\rangle = (1/n)e^{-i(n-1)\theta}P(e^{i\theta})$, which is real. We wish to maximize the real valued function $P(z)/z^{n-1}$, $z = e^{i\theta}$. Differentiating, we see that the maximum occurs when $zP'(z) - (n-1)P(z) = 0$, $|z| = 1$. Using these results, the following theorem gives a theoretical solution to the rank-1 problem. However, the computation of the roots of a polynomial is ill conditioned (with perturbation of the coefficients) and not practical for polynomials of large degree. In § 4, we propose a method based on equally spaced approximations, which in all our numerical examples yielded excellent initial starting values, so that a solution near the optimum value was obtained with minimal computation. This approach avoids the problem of computing all of the roots of the polynomial, and the method is also useful to find higher rank approximations.

THEOREM 4. *Let $A = [\alpha(j-k)]$, $\overline{\alpha(l)} = \alpha(-l)$, be a given matrix in $\mathscr{T}_n$, $A \geqq 0$, and let $P(z)$ be given by (6). Then $e^{-i(n-1)\theta}P(e^{i\theta}) \geqq 0$ for $0 \leqq \theta \leqq 2\pi$ and*

$$\|A - T\|^2 = \min_{S \in T_n^1} \|A - S\|^2 = \|A\|^2 - \frac{1}{n^2}|P(z)|^2,$$

where $z = e^{i\theta}$ is chosen to maximize $e^{-i(n-1)\theta}P(e^{i\theta})$ and $T = \langle A, T_0(\theta)\rangle T_0(\theta) = (1/n)e^{-i(n-1)\theta}P(e^{i\theta})T_0(\theta)$.

Proof. From the previous discussion, everything is clear except the fact that $e^{-i(n-1)\theta}P(e^{i\theta}) = n\langle A, T_0(\theta)\rangle$ is nonnegative. To see that this is true, we use the orthonormal basis given by Corollary 1. Write

$$(7) \qquad A = \sum_{l=0}^{n-1} a_l T_l(\theta) + \sum_{1 \leqq j < n/2} b_j R_j(\theta) + \sum_{1 \leqq k \leqq n/2} c_k S_k(\theta).$$

Let U be the first row of $T_0(\theta)$. Since A is positive semidefinite, $0 \leqq UAU^*$. However, $T_0(\theta) = nU^*U$ so using the orthogonality properties,

$$0 \leqq UAU^* = \langle A, U^*U\rangle = \frac{1}{n}\langle A, T_0(\theta)\rangle = \frac{1}{n} a_0.$$

Example 1. The problem of finding the best rank-1 approximation to a matrix A in $\mathcal{T}_n$ may not have a unique solution. Suppose A has the form

$$A = \sum_{j=0}^{n-1} c_j T_j(0).$$

As we have just noted, to find the best rank-1 approximation to A, we need to find

$$\max_\theta \langle A, T(\theta)\rangle = \max_\theta \sum_{j=0}^{n-1} c_j \langle T_j(0), T(\theta)\rangle$$

$$= \max_\theta \sum_{j=0}^{n-1} c_j \cdot \frac{1}{n} \frac{1 - \cos n\theta}{1 - \cos(\theta - 2j\pi/n)}.$$

However, from the proof of Lemma 3, we see that

$$\sum_{j=0}^{n-1} \frac{1 - \cos n\theta}{1 - \cos(\theta - 2j\pi/n)} = \sum_{j=0}^{n-1} \sum_{l,k=1}^{n} e^{-i(l-k)(\theta - 2j\pi/n)}$$

$$= \sum_{l,k=1}^{n} \sum_{j=0}^{n-1} e^{-i(l-k)(\theta - 2j\pi/n)}.$$

If $l \neq k$, the summation on j is zero. So the sum has the value $\sum_{l=1}^{n} \sum_{j=0}^{n-1} 1 = n^2$. Thus we have shown $\langle A, T(\theta)\rangle = \sum_{j=0}^{n-1} c_j \lambda_j / n$ where each $\lambda_j \geqq 0$ and $\sum_{j=0}^{n-1} \lambda_j = n^2$.

Therefore, $|\langle A, T(\theta)\rangle| \leqq \max_j c_j n$ and equality clearly holds when $\theta = 2l\pi/n$, $\max_j c_j = c_l$. The number of solutions to the minimum problem will therefore be the number of l such that $c_l = \max_j c_j$.

4. Higher rank approximations. Suppose $B \in \mathcal{T}_n^r$ so that $B = VDV^*$ where V is an $n \times r$ Vandermonde matrix and D is an $r \times r$ diagonal matrix with positive elements on the diagonal. Thus B is Hermitian, Toeplitz, and positive semidefinite of rank r. Now assume B and the rank r are unknown but measurements can be obtained to form an $n \times n$ matrix A that is a perturbation of B. We wish to use A to approximate r and B. If A is not Hermitian and Toeplitz, we replace A by $(\hat{A} + \hat{A}^*)/2$ (given by Theorem 1), and we rename this matrix A. Our goal is to find r, $t_1, t_2, \ldots, t_r$ and $\theta_1, \theta_2, \ldots, \theta_r$ so that $\sum_{l=1}^{r} t_l T(\theta_l)$ is B, where $T(\theta_l) = nT_0(\theta_l) = (e^{i(j-k)\theta_l})$. One serious problem in attempting to optimize $\|A - \sum_{l=1}^{r} t_l T(\theta_l)\|^2$ over the variables t_l, θ_l, $1 \leqq l \leqq r$ for fixed r, is that there are many local maxima, minima, and saddle points. Thus any numerical procedure for carrying out this optimization procedure requires that good starting values be obtained. Note that

$$\begin{aligned}\left\|A-\sum_{l=1}^{r} t_lT(\theta_l)\right\|^2 &= \|A\|^2-2\sum_{l=1}^{r} t_l\langle A, T(\theta_l)\rangle \\ &\quad +2\sum_{l=1}^{r-1}\sum_{p=l+1}^{r} t_lt_p\langle T(\theta_l), T(\theta_p)\rangle+\sum_{l=1}^{r} t_l^2\|T(\theta_l)\|^2 \\ &= \|A\|^2-2\sum_{l=1}^{r} t_le^{-i(n-1)\theta_l}P(e^{i\theta_l}) \\ &\quad +2\sum_{l=1}^{r-1}\sum_{p=l+1}^{r} t_lt_p\frac{1-\cos n(\theta_l-\theta_p)}{1-\cos(\theta_l-\theta_p)}+n^2\sum_{l=1}^{r} t_l^2.\end{aligned} \tag{8}$$

We proceed as follows. First approximate A by using n rank-1 matrices $T(\theta_l)$ with $\theta_1=\theta$, $\theta_2=\theta+2\pi/n, \ldots, \theta_n=\theta+2(n-1)\pi/n$ (equally spaced values). Actually, we find the projection of A on the space spanned by $\{T_l(\theta)\colon 1\leqq l\leqq n\}$, where $T_l(\theta)$ is defined by (3) and (5), and then we optimize over θ. This turns out to be surprisingly easy to do in closed form. Thus in (8) the terms involving $\langle T(\theta_l), T(\theta_p)\rangle$ are zero when $l\neq p$ and $t_l = 1/n^2\langle A, T(\theta_l)\rangle = 1/n^2e^{-i(n-1)\theta_l}P(e^{i\theta_l}) = 1/n^2e^{-i(n-1)(\theta+2(l-1)\pi/n)}P(e^{i(\theta+2(l-1)\pi/n)})$, where P is given by (6). Then (8) becomes

$$\left\|A-\sum_{l=1}^{n} t_lT(\theta_l)\right\|^2 = \|A\|^2-n^2\sum_{l=1}^{n} t_l^2, \tag{9}$$

where each t_l depends on θ as above. Let $Q(z)=\sum_{j=0}^{n-1} e^{4ij\pi/n}P^2(ze^{2ij\pi/n})$ and note that

$$\begin{aligned}Q(ze^{i2\pi/n}) &= \sum_{j=0}^{n-1} e^{4ij\pi/n}P^2(ze^{i2(j+1)\pi/n}) \\ &= e^{-4i\pi/n}\sum_{j=0}^{n-1} e^{4i(j+1)\pi/n}P^2(ze^{i2(j+1)\pi/n}) \\ &= e^{-4i\pi/n}Q(z).\end{aligned}$$

Since $\deg Q(z)\leqq 2(2n-2)=4(n-1)$, we have

$$Q(z)=\sum_{k=0}^{4(n-1)} q_kz^k.$$

However,

$$Q(ze^{i2\pi/n})=\sum_{k=0}^{4(n-1)} q_ke^{i2k\pi/n}z^k = e^{-4i\pi/n}\sum_{k=0}^{4(n-1)} q_kz^k;$$

thus $q_ke^{i2k\pi/n}=q_ke^{-4i\pi/n}$ and $q_k=q_ke^{2i(k+2)\pi/n}$. Hence $q_k=0$ except when $k+2=nl$ for some positive integer l and so q_{n-2}, q_{2n-2}, and q_{3n-2} are the only nonzero coefficients of $Q(z)$.

So $Q(z)=z^{n-2}(q_{n-2}+q_{2n-2}z^n+q_{3n-2}z^{2n})$, and from the definition of P it follows that $e^{-2i(n-1)\theta}Q(e^{i\theta})$ is real. Hence q_{2n-2} is real and $q_{3n-2}=\bar{q}_{n-2}$. Thus we want to maximize

$$q_{2n-2}+\bar{q}_{n-2}e^{in\theta}+q_{n-2}e^{-in\theta}.$$

Clearly, we want $n\theta = \arg(q_{n-2})$ (if $q_{n-2} = 0$, then (9) is a constant). Thus we have found θ to minimize (9). From the definition of Q, it is easy to check that $q_{n-2} = \sum_{l=1}^{n-1} l(n-l)\alpha(l)\alpha(n-l)$.

Now denote the right-hand side of (8) by $F(\theta_1, \theta_2, \dots, \theta_r, t_1, t_2, \dots, t_r)$. A necessary condition that F be a minimum is that $\partial F/\partial\theta_j = 0$ and $\partial F/\partial t_j = 0$, $1 \leqq j \leqq r$. We obtain starting values by using the r largest values of the t_l and the corresponding θ_l obtained above. Newton's method for $2r$ variables is then used to solve the $2r$ equations $\partial F/\partial\theta_j = 0$ and $\partial F/\partial t_j = 0$, $1 \leqq j \leqq r$. In practice, to obtain convergence to the best rank-r approximation, unless r is quite small, it is necessary to obtain the result in steps. In other words, find the best rank-r_1 approximation $T^{(1)}$ where r_1 is small, and set $A^{(1)} = A - T^{(1)}$. Find new starting values as described above for approximating $A^{(1)}$, and use these together with the values that give $T^{(1)}$ to obtain a rank-r_2 approximation $T^{(2)}$ where $r_2 > r_1$. Continue this process until the rank-r approximation is obtained.

5. Numerical results. Some numerical results are given in Tables 1–5. As yet, we do not have a proof of convergence to the desired optimum rank-r matrix, but the method has succeeded in determining the correct rank and closely approximates the weights and frequencies in a perturbed matrix as described below in all cases when we have increased the rank by 1 at each stage.

The results were obtained by applying the method of § 3 as follows. A matrix was formed from (1) by randomly choosing r weights d_{jj} in matrix D, $0 \leqq d_{jj} \leqq 4$, and r values θ_j, $0 \leqq \theta_j \leqq 2\pi$, to determine the Vandermonde matrix V. Typically n was chosen to be 60 and $r = 10$, although some other results are given. The matrix thus obtained by (1) was perturbed by adding random noise of the form $a + bi$, where $-1 \leqq a \leqq 1$, $-1 \leqq b \leqq 1$ to each entry of the matrix. The Hermitian Toeplitz matrix A is the matrix obtained by (2) from the matrix described above. The problem is to recover the r frequencies and weights that determine the matrix before the noise is added. The convergence criterion is that the maximum change of any of the variables be less than 5×10^{-5} (using Newton's method). The integer M is the number of iterations required for convergence. The method works best for relatively large n, as is to be expected. For example, with $n = 3$, $r = 2$, the noise seems to be too large to recover the original weights and frequencies, while with larger n, forming the Hermitian Toeplitz matrix using (2) tends to make the noise less significant.

Table 1 illustrates an example of the approximation described in § 3. The first two columns give the weights d_j and frequencies θ_j used to generate the matrix A before the noise is added (using the representation (1)). The matrix is 60×60 and of rank 10 before the perturbation. In the last five columns, the approximations are obtained, increasing the rank of the approximation by 1 at each step. L is the rank of the approximation, M is the number of iterations required to get convergence, and t_j and ϕ_j are the weights and frequencies in the approximating matrix. The last column gives the square of the norm of $A - T$ where T is the approximating matrix. The weights and frequencies are not listed for some of the approximations (i.e., ranks 3, 4, 6–9, 11) because the values do not change very much in successive approximations. Note that the norm of $A - T$ decreases rather dramatically as the rank of the approximation increases until the correct rank (10) is reached. Also, the additional weights beyond the rank-10 approximations are an order of magnitude smaller. Of course, the norm of $A - T$ is not zero when T has rank 10 because of the "noise" that remains.

In Table 2 we show the results of beginning with a higher rank approximation in the first step and increasing the rank by more than 1 in succeeding steps. In this particular case, we begin with a rank-4 approximation using initial values obtained from the ap-

TABLE 1

$N = 60$	$r = 10$
d_j	θ_j
3.66180	2.65965
3.29453	5.40062
2.73838	0.84639
2.35246	5.61409
1.98142	1.95922
1.85204	5.91687
1.16184	3.33427
0.59352	3.05046
0.34123	5.08016
0.29295	1.25646

$\|A\|^2 = 166699$

L	M	t_j	ϕ_j	$\|A - T\|^2$
1	4	3.65707	2.65974	118550
2	3	3.65683	2.65976	79730
		3.28388	5.40025	
3	4			59917
4	4			33254
5	3	3.65251	2.65981	19127
		3.28231	5.40051	
		2.34579	5.61464	
		2.72215	0.84623	
		1.98100	1.95921	
6	3			6976
7	3			2059
8	3			759
9	3			321
10	2	3.64690	2.65976	29
		3.28143	5.40042	
		2.34243	5.61429	
		2.72032	0.84623	
		1.97954	1.95924	
		1.83690	5.91673	
		1.16349	3.33485	
		0.60104	3.05017	
		0.34874	5.07902	
		0.28682	1.25872	
11	4			28
12	3	3.64686	2.65976	26
		3.28142	5.40042	
		2.34242	5.61429	
		2.72032	0.84623	
		1.97907	1.95926	
		1.83689	5.91673	
		1.16348	3.33485	
		0.60104	3.05017	
		0.34874	5.07902	
		0.28648	1.25907	
		0.02215	1.77852	
		0.01843	1.16700	

proximation by equally spaced frequencies as discussed in § 3. Subsequently, we increase the rank by 2 in successive steps. The matrix A is the same as in Table 1.

Our attempts to start with higher rank approximations in the first step indicate that for 60 × 60 matrices of rank 10 perturbed with random noise, if the initial approximation is by a matrix of rank 5 or more, then we get convergence to some critical value that is not the best approximation of that rank. However, in every case with various sizes of matrices, if we start with rank 1 and increase the rank by 1 at each step, there is a point at which the weight associated with the last frequency decreases dramatically as compared

TABLE 2

$N = 60$	$r = 10$					
d_j	θ_j	L	M	t_j	ϕ_j	$\|A - T\|^2$
3.66180	2.65965	4	4	3.65652	2.65977	47733
3.29453	5.40062			3.28248	5.40043	
2.73838	0.84639			2.34327	5.61429	
2.35426	5.61409			1.83975	5.91674	
1.98142	1.95922					
1.85204	5.91687	6	5	3.65247	2.65982	6976
1.16184	3.33427			3.28209	5.40044	
0.59352	3.05046			2.34299	5.61430	
0.34123	5.08016			1.83945	5.91675	
0.29295	1.25646			2.72204	0.84624	
				1.98087	1.95921	
$\|A\|^2 = 166699$						
		8	5	3.64707	2.65976	759
				3.28164	5.40042	
				2.34265	5.61431	
				1.83904	5.91675	
				2.72188	0.84623	
				1.98014	1.95923	
				1.16365	3.33485	
				0.60117	3.05017	
		10	6	3.64706	2.65976	320
				3.28147	5.40042	
				2.34247	5.61429	
				1.83905	5.91675	
				2.72173	0.84623	
				1.98009	1.95923	
				1.16352	3.33485	
				0.60105	3.05018	
				0.34880	5.07902	
				0.01684	1.14863	
		12	3	3.64687	2.65976	27
				3.28142	5.40042	
				2.34243	5.61429	
				1.83902	5.91675	
				2.72168	0.84622	
				1.97951	1.95925	
				1.16349	3.33485	
				0.60104	3.05017	
				0.34874	5.07902	
				0.01838	1.16658	
				0.28650	1.25904	
				0.02270	1.56761	

to the other weights, and the change in the norm of $A - T$ is quite small. This seems to be a clear indication that the appropriate rank has been found. These conclusions are based purely on experimental evidence and not on theoretical conclusions at this time.

Table 3 gives an example that shows that even though two of the frequencies are quite close, θ_3 and θ_9, our method eventually identified the existence of these two frequencies. Starting with a rank-4 approximation and increasing the rank by 2 at each step, up through the rank-8 approximation, a frequency intermediate between the two near frequencies was obtained with a weight significantly larger than either of the associated

TABLE 3

$N = 60$	$r = 10$					
d_j	θ_j	L	M	t_j	ϕ_j	$\|A - T\|^2$
3.51524	6.19195	4	4	4.66183	5.89391	129217
3.46610	5.15524			3.47806	5.15559	
3.38017	5.87692			3.62805	6.19065	
3.31638	1.71590			3.39946	1.71666	
3.03087	5.36127					
2.93608	0.04048	6	3	4.66093	5.89382	66693
2.84050	3.57764			3.47721	5.15534	
2.53801	1.34266			3.62788	6.19068	
2.16784	5.92609			3.39863	1.71663	
1.50798	1.88239			3.04486	5.36160	
				2.84550	3.57755	
$\|A\|^2 = 340389$						
		8	5	4.66062	5.89382	12148
				3.47668	5.15534	
				3.62770	6.19068	
				3.37924	1.71682	
				3.04440	5.36159	
				2.84447	3.57755	
				3.08523	0.04323	
				2.56000	1.34226	
		10	13	3.35669	5.87631	384
				3.47218	5.15525	
				3.60788	6.19015	
				3.32175	1.71603	
				3.03762	5.36137	
				2.84378	3.57751	
				3.08042	0.04344	
				2.55852	1.34236	
				1.51771	1.88250	
				2.19744	5.92577	

weights. The rank-10 approximation then successfully identified the existence of these two frequencies. As has been typical in our computations, the fact that these frequencies were so close forced more iterations than usual (13) to obtain convergence.

TABLE 4

$N = 10$	$r = 3$					
d_j	θ_j	L	M	t_j	ϕ_j	$\|A - T\|^2$
3.68290	4.24646	1	4	4.12979	4.13429	867
2.79552	3.76466					
1.56956	0.64161	3	5	3.70186	4.25141	6
				2.83512	3.77076	
$\|A\|^2 = 2572$				1.48886	0.64123	
		5	3	3.67682	4.25378	3
				2.82468	3.77694	
				1.48593	0.64177	
				0.13880	2.17547	
				0.07060	3.41188	

TABLE 5

$N = 4$	$r = 2$					
d_j	θ_j	L	M	t_j	ϕ_j	$\Vert A - T\Vert^2$
2.39610	5.15882	1	2	3.52520	5.04331	1
1.28269	4.75756					
		2	5	3.52517	5.04346	0.6
$\Vert A\Vert^2 = 200$				0.16967	0.37015	

Since the process of averaging the diagonals to obtain a Toeplitz matrix tends to minimize the effect of the noise, it is to be expected that the noise would be more significant in smaller matrices. That is indeed the case. Our computations have shown that for matrices as small as 10 × 10, the results remain quite good (especially for identifying frequencies). The results are not as good in the 4 × 4 case. See Tables 4 and 5.

6. Unrestricted rank case. In this section we consider the special case of Problem 1 when $r = n - 1$ and D has a negative eigenvalue, and solve

$$(10) \qquad \min_{T \in \mathcal{T}_n^{n-1}} \Vert D - T\Vert .$$

The basic idea is to view $\mathcal{T}_n^n$ as the intersection of the convex cone K of positive semidefinite $n \times n$ Hermitian matrices and the subspace $\mathcal{T}_n$ of all $n \times n$ Hermitian Toeplitz matrices. Hence $\mathcal{T}_n^{n-1}$ is a boundary of $\mathcal{T}_n^n$ and the unique solution to (10) can be obtained by a modified alternating projection algorithm. Von Neuman showed that if S_1 and S_2 are subspaces and D is a given point, then the nearest point to D in $S_1 \cap S_2$ could be obtained by the following algorithm.

```
ALTERNATING PROJECTION ALGORITHM.
    Let $X_1 = D$
        For $k = 1, 2, \cdots$
            $X_{k+1} = P_1(P_2(X_k))$.
```

X_k converges to the near point D in $S_1 \cap S_2$ where P_1 and P_2 are the orthogonal projections on S_1 and S_2, respectively. The necessary modification of von Neumann's algorithm when S_1 and S_2 are replaced by convex sets was first given by Boyle and Dykstra [2]. Other proofs and connections to duality along with applications were given by Han [11] and Gaffke and Mathar [6]. These modifications were applied in [6] and [7] to find the nearest distance matrix to a given data matrix. Since the constraint set for the distance matrix problem consists of the intersection of a convex cone and a subspace, the ideas of [7] carry over directly to the solution of (10). The necessary modification of von Neumann's algorithm, as given in [2], [7], and [11] when applied to our setting, yields the following modified alternating projection algorithm MAPT (see [7, eq. (4.4)]). Given a Hermitian data matrix D, which is not positive semidefinite, we get the following algorithm.

```
ALGORITHM MAPT.
    Let $F_1 = D$
        For $j = 1, 2, 3 \cdots$
            $F_{j+1} = F_j + [P_T P_K(F_j) - P_K(F_j)]$.
```

Then $\{P_K(F_k)\}$ and $\{P_T P_K(F_k)\}$ converge in a Frobenius norm to the solution of (8). Here $P_K(F)$ is the projection (proximity map) of F onto the convex cone K of positive semidefinite Hermitian matrices. One finds $P_K(F)$ (see [12]) by finding a spectral decomposition of F and setting the negative eigenvalues to zero. Finding the spectral decomposition is $O(n^3)$ work, and hence $P_K(F)$ is the major computational work of the algorithm. $P_T(F)$ is the projection onto the subspace of Hermitian Toeplitz matrices $\mathscr{T}_n$. Here one thinks of $\mathscr{T}_n$ as complex matrices, but scalar multiplication is only over the reals. $P_T(F)$ is easy to compute using Lemma 2. Simply set each diagonal in $P_T(F)$ to be the average of the corresponding diagonal of F.

An application of Algorithm MAPT may be used to solve the following matrix minimization problem.

PROBLEM S. Given a positive definite Hermitian matrix D, find the nearest positive semidefinite (positive definite) Hermitian Toeplitz matrix T to D in the Frobenius norm.

Solution. Step 1. Find the nearest Toeplitz matrix $\hat{T}$ to D by averaging the diagonals. If $\hat{T}$ is positive definite (or positive semidefinite), then one is finished.

Step 2. If $\hat{T}$ is not positive semidefinite, then use Algorithm MAPT to find the nearest positive semidefinite Toeplitz matrix T to $\hat{T}$.

Now T is the solution by an argument similar to the proof of Theorem 2 because $\hat{T}$ was obtained by a projection, and hence for every Toeplitz matrix S, $\|D - S\|^2 = \|D - \hat{T}\|^2 + \|\hat{T} - S\|^2$. Since our Algorithm MAPT minimized the last term, we are done. The rate of convergence of Algorithm MAPT is linear, and numerical results for the related case of approximation of distance matrices by modified alternating projections may be found in [7].

REFERENCES

[1] G. AMMAR, W. GRAGG, AND L. REICHEL, *Determination of Pisarenko frequency estimates as eigenvalues of an orthogonal matrix*, in Advanced Algorithms and Architectures for Signal Processing II, Vol. 826, F. Luk, ed., Proc. Society of Photo-Optical Instrumentation Engineers, The Internat. Society for Optical Engineering, Bellingham, WA, 1987, pp. 143–145.

[2] J. P. BOYLE AND R. L. DYKSTRA, *A method for finding projections onto the intersection of convex sets in Hilbert space*, Lecture Notes in Statistics 37, Springer-Verlag, Berlin, 1986, pp. 28–47.

[3] J. P. BURG, D. G. LUENBERGER, AND D. L. WENGER, *Estimation of structured covariance matrices*, Proc. IEEE, 70 (1982), pp. 963–974.

[4] G. CYBENKO, *Moment problems and low rank Toeplitz approximations*, Circuits Systems Signal Process., 1 (1982), pp. 345–365.

[5] K. FAN AND A. HOFFMAN, *Some metric inequalities in the space of matrices*. Proc. Amer. Math Soc., 6 (1955), pp. 111-116.

[6] N. GAFFKE AND R. MATHAR, *A cyclic projection algorithm via duality*, Metrika, 36 (1989), pp. 29–54.

[7] W. GLUNT, T. HAYDEN, S. HONG, AND J. WELLS, *An alternating projections algorithm for computing the nearest Euclidean distance matrix*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 589-600.

[8] W. GLUNT, T. L. HAYDEN, AND W. LIU, *The embedding problem for predistance matrices,* Bull. Math. Biol., 53 (1991), pp. 769–796.

[9] W. B. GRAGG, *Positive definite Toeplitz matrices, the Arnoldi process for isometric operators and Gaussian quadrature on the unit circle*, in Numerical Methods in Linear Algebra, E. S. Nikolaev, ed., Moscow University Press, Moscow, 1982, pp. 16–32. (In Russian.)

[10] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, CA, 1958.

[11] S. HAN, *A successive projection method*, Math. Programming, 40 (1988), pp. 1–14.

[12] N. J. HIGHAM, *Computing a nearest symmetric positive semi-definite matrix*, Linear Algebra Appl., 103 (1988), pp. 103–118.

[13] S. Y. KUNG, *A Toeplitz approximation method and some applications*, Internat. Sympos. on Mathematical Theory of Networks and Systems, Vol. IV, Western Periodicals Co., North Hollywood, CA, 1981, pp. 262–266.

[14] L. REICHEL AND G. S. AMMAR, *Fast approximations of dominant harmonics by solving an orthogonal eigenvalue problem*, in Signal Processing II, J. McWhirter, ed., Oxford University Press, Oxford, 1990, pp. 575–591.

[15] A. K. SHAW AND R. KUMARESAN, *Some structured matrix approximation problems*, Proc. IEEE ICASSP, Vol. 4, Internat. Conf. Acoustics, Speech, and Signal Processing, 1988, pp. 2324–2327.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 735–765, July 1993

A WEAKLY STABLE ALGORITHM FOR PADÉ APPROXIMANTS AND THE INVERSION OF HANKEL MATRICES*

STAN CABAY† AND RON MELESHKO‡

Abstract. A new algorithm, Algorithm NPADE, is presented for numerically computing Padé approximants in a weakly stable fashion. By this, it is meant that if the problem is well conditioned, then Algorithm NPADE produces a good solution. No restrictions are imposed on the problem being solved. Except in certain pathological cases, the cost of the algorithm is $O(n^2)$, where n is the maximum degree of the polynomials comprising the Padé approximant.

The operation of Algorithm NPADE is controlled by a single parameter. Bounds are obtained for the computed solution and it is seen that they are a function of this parameter. Experimental results show that the bounds, while crude, reflect the actual behavior of the error. In addition, it is shown how better bounds can easily be obtained a posteriori. As another application of Algorithm NPADE, it is shown that it can be used to compute stably, in a weak sense, the inverse of a Hankel or Toeplitz matrix.

Key words. Padé approximants, numerical computation, Hankel inverses, Toeplitz inverses, numerical stability

AMS subject classifications. 41A21, 65F05, 65G05

1. Introduction.

Let

$$A(z) = \sum_{i=0}^{\infty} a_i z^i \quad \text{and} \quad B(z) = \sum_{i=0}^{\infty} b_i z^i$$

be formal power series, where $a_i, b_i \epsilon \Re$. We consider the problem of approximating $A(z)/B(z)$ and assume, without loss of generality, that $b_0 \neq 0$.[1] One way to approximate such a power series quotient is to use rational functions and a method dating back to the last century (Padé [49], Frobenius [27]).

DEFINITION 1.1. Let

$$U(z) = \sum_{i=0}^{m} u_i z^i \quad \text{and} \quad V(z) = \sum_{i=0}^{n} v_i z^i$$

be polynomials of degrees m and n, respectively. $(U(z), V(z))$ is said to be a *Padé form of type* (m, n) for the power series quotient $A(z)/B(z)$ if $V(z)$ is nonzero and

$$A(z)V(z) + B(z)U(z) = z^{m+n+1}W(z). \tag{1}$$

$W(z)$ is called the *residual* for the (m, n) Padé form.

This definition is more general than that which normally is used (cf. Gragg [31]). In practice, the case where $B(z) \equiv -1$ is most often considered. For the moment, we let $B(z) \equiv -1$, but later we will often revert to the more general formulation.[2]

* Received by the editors December 19, 1990; accepted for publication (in revised form) October 31, 1991.

† Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2H1 (cabay@cs.ualberta.ca). The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A8035.

‡ Department of Biochemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2H7 (ron@odysseus.biochem.ualberta.ca).

[1] If $b_0 = 0$ and $a_0 \neq 0$, then $A(z)$ and $B(z)$ can be interchanged. If both are zero, then if b_λ (or a_λ) is the first nonzero coefficient, we can work with $z^{-\lambda}A(z)$ and $z^{-\lambda}B(z)$.

[2] This more general formulation simplifies the notation used in this paper and is also in keeping with the formulation of Padé–Hermite approximants.

A common way to organize the Padé approximants for a fixed A/B is given by the following.

DEFINITION 1.2. The *Padé table* for $A(z)/B(z)$ is the doubly infinite array whose (m, n) entry is a Padé form of type (m, n) for A/B.

Padé approximants are useful in many diverse fields of science. For example, Baker and Graves-Morris [4], [5] survey both the theory and applications of Padé approximants, while Bultheel and Van Barel [12] provide an extensive survey of Padé techniques in system theory. Gragg [31] gives an introduction to the Padé fraction in the context of the algorithms of numerical analysis.

Padé approximants can also be viewed as a matrix problem. From (1), with $B(z) = -1$, we obtain

$$\begin{pmatrix} a_0 & & 0 & -1 & & 0 \\ & \ddots & & 0 & \ddots & \\ \vdots & & a_0 & & & -1 \\ & & \vdots & \vdots & & \vdots \\ a_{m+n} & & a_m & 0 & & 0 \end{pmatrix} \begin{pmatrix} v_0 \\ \vdots \\ v_n \\ u_0 \\ \vdots \\ u_m \end{pmatrix} = 0.$$

This homogeneous system is underdetermined and so has a nontrivial solution. Letting v_0 be arbitrary, we can decouple the system to get

$$(2) \qquad \begin{pmatrix} a_m & \cdots & a_{m-n+1} \\ \vdots & & \vdots \\ a_{m+n-1} & \cdots & a_m \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = -v_0 \begin{pmatrix} a_{m+1} \\ \vdots \\ a_{m+n} \end{pmatrix}$$

and

$$(3) \qquad \begin{pmatrix} u_0 \\ \vdots \\ u_m \end{pmatrix} = \begin{pmatrix} a_0 & & 0 \\ & \ddots & \\ \vdots & & a_0 \\ & & \vdots \\ a_m & \cdots & a_{m-n} \end{pmatrix} \begin{pmatrix} v_0 \\ \vdots \\ v_n \end{pmatrix}.$$

$V(z)$ can be obtained from (2) and then $U(z)$ is easily obtained from (3). The matrix on the left-hand side of equation (2) is a Toeplitz matrix. Now the order of the variables in (2) is arbitrary; we could write it as

$$(4) \qquad \begin{pmatrix} a_{m-n+1} & \cdots & a_m \\ \vdots & & \vdots \\ a_m & \cdots & a_{m+n-1} \end{pmatrix} \begin{pmatrix} v_n \\ \vdots \\ v_1 \end{pmatrix} = -v_0 \begin{pmatrix} a_{m+1} \\ \vdots \\ a_{m+n} \end{pmatrix}.$$

In this case, the matrix on the left-hand side is a Hankel matrix. We shall denote it by $H_{m,n}$.

It is clear that finding a solution to the Toeplitz or Hankel system defined by equations (2) or (4) is equivalent to solving the Padé approximant problem. If the Toeplitz system in (2) is symmetric, then (2) is known as the Yule–Walker equation. Such systems arise in the analysis of sunspot data (Yule [63]) and in time series analysis (Walker [58]). Similar series arise in signal processing (Levinson [40]), linear prediction (Makhoul [44]), and linear filtering (Kailath [37]).

The general case with an arbitrary $A(z)$ and $B(z)$ has a similar formulation. Letting v_0 be arbitrary, (1) gives us

$$(5)\qquad \begin{pmatrix} 0 & & 0 & b_0 & & 0 \\ a_0 & & \vdots & \vdots & \ddots & \vdots \\ & \ddots & & & & b_0 \\ \vdots & & a_0 & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ & & \vdots & & & \vdots \\ a_{m+n-1} & \cdots & a_m & b_{m+n} & \cdots & b_n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \\ u_0 \\ \vdots \\ u_m \end{pmatrix} = -v_0 \begin{pmatrix} a_0 \\ \vdots \\ \vdots \\ a_{m+n} \end{pmatrix}.$$

Again, the system has a nontrivial solution with $V(z) \neq 0$. The matrix on the left-hand side of (5) is known as a Sylvester matrix; we will denote it by $S_{m,n}$.

2. Numerical considerations and stability. Due to the regular structure of systems (2) and (4), it is possible to solve them with methods that use less than $O(n^3)$ operations (as would be required if Gaussian elimination were used). One of the earliest such methods is that of Levinson [40], which dates back to 1947. He solved the problem for symmetric Toeplitz systems, though he was not explicit in his use of matrix methods. His algorithm has a cost complexity of $O(n^2)$. Algorithms for solving Toeplitz or Hankel systems with this time complexity are referred to as *fast* algorithms (Ammar and Gragg [1]). Levinson's algorithm was subsequently extended to the nonsymmetric case (Trench [55], Zohar [64], [65]); however, all these methods required the Toeplitz matrix and its principal minors to be nonsingular. Rissanen [50] developed an algorithm for Hankel matrices that overcame the restriction of nonsingularity of the principal submatrices. Other fast methods have appeared (Bareiss [6], Jain [36], Gragg et al. [32], Delsarte and Genin [25], Cabay and Choi [16], Bultheel [11]). Recently, *superfast* $O(n \log^2 n)$ algorithms have appeared (Bitmead and Anderson [9], Brent, Gustavson, and Yun [10], Cabay and Choi [16], de Hoog [24], Ammar and Gragg [1]). Both the fast and superfast algorithms rely upon the solution of smaller subproblems. They are expressed in terms of submatrices or bordering techniques and are equivalent to Gaussian elimination without pivoting.

Most of these algorithms, especially those dealing with singular principal submatrices, are posed in an algebraic setting, in the sense that they either explicitly assume exact arithmetic or implicitly do so by not considering the effects of rounding errors. While there exist algebraic programming systems such as Maple, Mathematica, and Macsyma that enable the easy coding of these methods (Geddes [28], Cabay and Kossowski [18]), these algebraic algorithms are expensive and so they are limited, in practice, to small problems.[3] For large problems, which can arise in many areas (Jain [36]), the only recourse is to use numerical methods. In this case, the algebraic algorithms can be shown to be unstable for the general problem (Bultheel and Wuytack [13]). Some work has been done on the numerical aspects of their computation, but either it concentrates on numerical experiments (Luke [42], [43]) or else the analysis is only valid for restricted classes of problems (Wynn [62], Cybenko [23]). (For example,

[3] See, for example, the timings reported in Geddes [28].

Ammar and Gragg [2] present experimental results for a superfast solver applied to positive-definite symmetric Toeplitz matrices.) Some algorithms fail if singular subproblems are encountered (Claessens [21]), while other algorithms handle the singular cases by recognizing them and then using special rules (Rissanen [50], Claessens and Wuytack [22], and Bultheel [11]). But problems arise if the near-singular case is encountered. For example, Rissanen's algorithm requires the detection of a zero value, but if the value is nonzero but small, then, in the context of Gaussian elimination, what the algorithm does is equivalent to performing a column reduction with an extremely small pivot. This is known to be an unstable operation and poor solutions can result. Other algorithms fail for the same sort of reasons. (See, for example, Bultheel and Wuytack [13], Bunch [14], Luk and Qiao [41].)

The same problem of numerical instability also arises in the algorithms for scalar, vector, and matrix rational interpolation recently proposed by Antoulas et al. (see [3] and the references given there), by Beckermann [7], and by Van Barel and Bultheel (see [56], [57] for the scalar case). The approach of Van Barel and Bultheel (for the special case that all the interpolation points are at the origin) can be seen to be similar to the one of Cabay and Choi [16], for which we give a stabilized form here. Also in the case that the interpolation points are all at the origin (which is the Padé approximation problem), Beckermann, Cabay, and Labahn (see [38], [20], and [19]) generalize the Cabay and Choi algorithm to the vector and matrix case. These generalized algorithms appear also to have similar stable versions (see Cabay and Jones [17]). In addition, there appears to be a certain relationship [33] of these algorithms to those of Antoulas and Beckermann.

Bunch [14], [15] gives the best overview of the state of the art in a numerical setting; he provides an extensive survey of the limitations and shortcomings of many of the available methods for solving Toeplitz systems. If the matrix is positive-definite symmetric, the methods perform well numerically, but in the general case they can perform poorly. (Similar observations are made by Bultheel and Wuytack [13].) The observation that is made is that the positive-definite symmetric matrices are a class of matrices that can be solved stably using Gaussian elimination without pivoting. Bunch suggests that in order to make the algorithms stable for the general problem some pivoting must be introduced, but then this would result in the loss of the regular structure of the matrices that enables the algorithms to perform in a fast manner.

Sweet [53] takes a different approach using orthogonal transformations to produce a fast QR factorization of a Toeplitz matrix. Again, the regular structure of the Toeplitz matrix is exploited, but even this more stable technique has its problems. Luk and Qiao [41] give an example of a matrix for which Sweet's factorization produces poor results.

To be more precise, we need to define what we mean by stable. The following is a common definition of algorithmic stability (Bunch [15], Wilkinson [60]).

DEFINITION 2.1. An algorithm for solving linear equations is *stable*[4] for a class of matrices $\mathcal{M}$ if for each M in $\mathcal{M}$ and for each b the computed solution x_c to $Mx = b$ satisfies $\hat{M}x_c = \hat{b}$, where $\hat{M}$ is close to M and $\hat{b}$ is close to b.

However, for our purposes, we will use a weaker type of stability. Users often want to obtain solutions that are close to the true solution; they may not need (or want) to know that their solution is the exact solution of a nearby problem (as in Definition 2.1). As is well known, the final solution is dependent not only on the algorithm, but also upon the condition of the problem being solved. If the problem is poorly

[4] This type of stability is sometimes referred to as *backward stability*.

conditioned, then we cannot expect the computed solution to be good because just formulating the problem numerically can introduce perturbations. But if the problem is well conditioned, then we would like an accurate solution. To describe this, the following was introduced by Bunch [15].

DEFINITION 2.2. An algorithm for solving linear equations is *weakly stable*[5] for a class of matrices $\mathcal{M}$ if for each well-conditioned M in $\mathcal{M}$ and for each b, the computed solution x_c to $Mx = b$ is such that $\| x - x_c \|/\| x \|$ is small.

This type of algorithmic stability is not as strong as the commonly used type of stability. Stability implies weak stability, but not vice versa. In particular, weak stability says nothing about the behavior of the algorithm for ill-conditioned problems. On the other hand, it can be easier to show weak stability as the effects of the numerical operations do not need to be reflected back into the problem being solved.

The purpose for which Bunch introduced weak stability was to address the confusion about the stability of Toeplitz solvers. What he showed was that the algorithms he considered required that various submatrices be well conditioned. A submatrix can be ill conditioned even though the matrix itself is well conditioned. The well-conditioning of principal submatrices, however, can be guaranteed if the matrix is positive-definite symmetric. Thus, he concluded, the algorithms were weakly stable for the class of positive-definite symmetric matrices, but were not weakly stable for more general classes. For example, Cybenko [23], considered the stability of the Levinson–Durbin algorithm for the solution of Toeplitz systems. Cybenko showed that for the class of well-conditioned, positive-definite symmetric Toeplitz matrices the algorithm produced a solution with a small residual. Though he did not use the terminology, he had shown the weak stability of the algorithm for this class of Toeplitz matrices.[6]

This paper introduces a new algorithm, Algorithm NPADE, for the computation of Padé approximants. We will show that it is weakly stable for the general class of power series. Moreover, it is robust in the sense that it is capable of recognizing problems as they arise. In §3 some notation, basic definitions, and fundamental results are presented. These results in §3 are not new; rather, they are used to justify the validity of Algorithm NPADE which is presented in §4. The analysis of Algorithm NPADE follows, starting in §5 with a summary of the principal results. The details of the analysis are contained in §§6 and 7; these two sections contain a round-off analysis of the algorithm and are quite technical. The paper concludes with a brief discussion in §8 of some experimental results and in §9 with comments on further research.

However, before proceeding with this material, we complete this preliminary discussion of numerical matters, by now briefly reviewing, without proof, some standard results from the field of floating point error analysis. Lemmas 2.3, 2.6, 2.7, and 2.8 all refer to floating point numbers. We use μ to denote the unit round-off and assume that the degrees of all polynomials and the orders of all matrices are bounded by some M, where $M\mu \leq 0.01$ (this restriction comes from Forsythe and Moler [26]). After Wilkinson [59], we denote a floating point operation by $fl(\cdot)$.

LEMMA 2.3. *If* $n\mu \leq 0.01$, *then*

$$fl(\textstyle\sum_{k=1}^{n} u_k v_k) = \displaystyle\sum_{k=1}^{n} u_k v_k (1 + \gamma_k),$$

where $|\gamma_k| \leq 1.01 n\mu$.

[5] This type of stability is sometimes referred to as *forward stability*.

[6] Bunch [15] first observed that Cybenko's stability result was actually a proof of weak stability.

In our analysis, we almost always use polynomials; the following definition gives a polynomial norm that is defined in terms of vector norms. We also use power series, so the infinity norm for a power series is likewise defined.

DEFINITION 2.4. If $U(z) = \sum_{i=0}^{n} u_i z^i$, then define the infinity norm for $U(z)$ to be

$$\| U \|_\infty = \| (u_0, u_1, \ldots, u_n)^T \|_\infty .$$

If $A(z) = \sum_{i=0}^{\infty} a_i z^i$, then define the infinity norm for $A(z)$ to be

$$\| A \|_\infty = \sup\{|a_i| : 0 \le i < \infty\}.$$

DEFINITION 2.5. If $A(z) = \sum_{i=0}^{\infty} a_i z^i$, then $A \pmod{z^k} = \sum_{i=0}^{k-1} a_i z^i$.

LEMMA 2.6. *Let $A(z)$ be a power series and let $V(z)$ be a polynomial of degree n. Then*

$$\| (AV) \pmod{z^{m+1}} \|_\infty \le (\min\{m,n\}+1) \| A \|_\infty \| V \|_\infty .$$

LEMMA 2.7. *Let $A(z)$ be a power series and let $V(z)$ be a polynomial of degree n. Then*

$$fl((AV) \pmod{z^{m+1}}) = (AV) \pmod{z^{m+1}} + \Psi,$$

where

$$\| \Psi \|_\infty \le 1.01(\min\{m,n\}+1)^2 \| A \|_\infty \| V \|_\infty \mu.$$

LEMMA 2.8. *Let U and V be polynomials of degree n. Then*

$$fl(U+V) = U + V + \Psi,$$

where

$$\| \Psi \|_\infty \le \| U + V \|_\infty \mu.$$

Proofs of these lemmas can be found in Meleshko [46], [47]. As we use only the infinity norm in the following analysis, we simply use $\| \cdot \|$ instead of the more explicit $\| \cdot \|_\infty$.

3. Some basic definitions and notation. Cabay and Choi [16] present a polynomial algorithm that iteratively computes Padé forms along a fixed, but arbitrary, diagonal of the Padé table in an algebraic setting.[7] This computation along a fixed diagonal corresponds to solving Hankel systems of increasing order. It is equivalent to Rissanen's algorithm, but it has a lower operation count as it is better able to exploit the structure of the problem through the use of polynomial operations. Some of the results that form the basis of their algorithm are necessary to understand Algorithm NPADE as well. We present these results in a considerably different fashion; proofs can be found in Meleshko [46], [47] and the original proofs in [16]. We also assume that $m = n - 1$. It is easy to show that if we can obtain the $(n-1, n)$ Padé approximant, then the general (m, n) Padé approximant can also be obtained.

[7] Their algorithm is similar to ones described in Berlekamp [8] and Massey [45]. Moreover, in the case of a normal power series and computation along the superdiagonal of the Padé table, it is equivalent to the algorithm described in Gragg et al. [32].

With $A(z)$ and $B(z)$ fixed, we define a sequence of integers

$$n_0, n_1, n_2, \ldots,$$

by setting $n_0 = 0$ and $n_i = n_{i-1} + s_{i-1}$, where $s_{i-1} \geq 1$ is selected so that the Sylvester matrix $S_{n_i} = S_{n_i-1,n_i}$ is nonsingular. Because S_{n_i} is nonsingular, both the $(n_i - 2, n_i - 1)$ and $(n_i - 1, n_i)$ Padé forms can be made unique by imposing appropriate conditions (Gragg [31], Labahn and Cabay [38]). We use the following definition.

DEFINITION 3.1. The *normalized Padé approximant* of type (m, n) for the power series quotient $A(z)/B(z)$ is the (m, n) Padé form for which $\| U(z) \| + \| V(z) \| = 1$.

With $n_0, n_1, \ldots,$ as above, define $\mathcal{P}^{(i)}$ to be

$$\mathcal{P}^{(0)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$\mathcal{P}^{(i)} = \begin{pmatrix} P^{(i)} & Q^{(i)} \\ U^{(i)} & V^{(i)} \end{pmatrix}, \qquad i = 1, 2, \ldots,$$

where $(U^{(i)}, V^{(i)})$ is the $(n_i - 1, n_i)$ normalized Padé approximant for $A(z)/B(z)$ and $(P^{(i)}, Q^{(i)})$ is the $(n_i - 2, n_i - 1)$ normalized Padé approximant, multiplied by z^2, for $A(z)/B(z)$. Let

$$\mathcal{A} = \begin{pmatrix} B(z) \\ A(z) \end{pmatrix}$$

and denote the residual sequence associated with these Padé forms by

$$\{\mathcal{R}^{(i)}\} = \left\{ \begin{pmatrix} R^{(i)} \\ W^{(i)} \end{pmatrix} \right\}, \qquad i = 0, 1, \ldots.$$

Then $\mathcal{P}^{(i)}$ satisfies

$$\mathcal{P}^{(i)} \mathcal{A} = z^{2n_i} \mathcal{R}^{(i)}. \tag{6}$$

We refer to $\mathcal{P}^{(i)}$ as the n_ith Padé approximant pair for $\mathcal{A}$. (We could also define the Padé approximant pair for the case where S_{n_i} is singular; in this case, the Padé approximant pair would be nonunique.)

An important result is the following.

THEOREM 3.2.

$$S_{n_i} \textit{ is nonsingular } \Leftrightarrow r_0^{(i)} v_0^{(i)} \neq 0,$$

where $r_0^{(i)}$ *and* $v_0^{(i)}$ *are the constant terms in* $R^{(i)}$ *and* $V^{(i)}$, *respectively.*

The product $r_0^{(i)} v_0^{(i)}$ determines the algebraic property of nonsingularity of the Sylvester matrix; we will see shortly that the size of this quantity is also an indicator of the condition number of the matrix S_{n_i}.

We can view $W_i(z)/R_i(z)$ as a power series quotient and compute its normalized Padé approximants as well. Let $\hat{S}_{s_i}$ be the Sylvester system associated with the $(s_i - 1, s_i)$ Padé form for W_i/R_i. We define $\hat{\mathcal{P}}^{(i)}$ for W_i/R_i to be

$$\hat{\mathcal{P}}^{(i)} = \begin{pmatrix} 0 & z^2 \\ \hat{U}^{(i)} & \hat{V}^{(i)} \end{pmatrix} \quad \text{if } s_i = 1,$$

and

$$\hat{\mathcal{P}}^{(i)} = \begin{pmatrix} \hat{P}^{(i)} & \hat{Q}^{(i)} \\ \hat{U}^{(i)} & \hat{V}^{(i)} \end{pmatrix} \quad \text{for } s_i > 1,$$

where $(\hat{U}^{(i)}, \hat{V}^{(i)})$ is an $(s_i - 1, s_i)$ normalized Padé approximant for W_i/R_i and $(\hat{P}^{(i)}, \hat{Q}^{(i)})$ is an $(s_i - 2, s_i - 1)$ normalized Padé approximant, multiplied by z^2, for W_i/R_i. We refer to $\hat{\mathcal{P}}^{(i)}$ as the s_ith Padé approximant pair for $\mathcal{R}^{(i)}$. Note that

$$(7) \qquad \hat{\mathcal{P}}^{(i)}\mathcal{R}^{(i)} = O(z^{2s_i}).$$

From (6) and (7), it is easy to prove the following lemma.

LEMMA 3.3. *If $S_{n_i}$is nonsingular, then*

$$\bar{\mathcal{P}} = \hat{\mathcal{P}}^{(i)}\mathcal{P}^{(i)}$$

is an n_{i+1}th Padé approximant pair for $\mathcal{A}$.

Thus, by solving the associated residual system for a given Padé approximant, we can obtain a higher-order Padé form. However, if the residuals satisfy certain conditions, we can say more than this.

THEOREM 3.4. *If S_{n_i} is nonsingular, then*

$$\hat{S}_{s_i} \textit{ is nonsingular } \Leftrightarrow S_{n_{i+1}} \textit{ is nonsingular.}$$

Now Lemma 3.3 and Theorem 3.4 allow us to generate $\mathcal{P}^{(i)}$ iteratively.

COROLLARY 3.5. *If S_{n_i} is nonsingular, then, with α and β appropriately chosen nonzero scalar values,*

$$\mathcal{P}^{(i+1)} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} \hat{\mathcal{P}}^{(i)}\mathcal{P}^{(i)}$$

gives us the $(n_i + s_i - 1, n_i + s_i)$ normalized Padé approximant pair for $\mathcal{A}$.

These results define a three-term recurrence relation among the distinct entries along the superdiagonal of the Padé table. They are actually valid for any (m, n) and so define a three-term recurrence relation among the distinct entries along any diagonal of the Padé table.

The algorithm of Cabay and Choi selects s_i so that it is the first stepsize s_i for which $\hat{S}_{s_i}$ is nonsingular; this is easily determined if exact arithmetic is assumed. (From a matrix perspective, equivalently, Rissanen's algorithm selects the smallest s_i for which $S_{n_i+s_i}$ is nonsingular.) However, by using a different criteria for the selection of the stepsize, a numerical algorithm that is weakly stable is derived; it is given below in §4 with an analysis in the succeeding sections. This criteria will turn out to be a measure of the conditioning of the Hankel system corresponding to the currently computed Padé approximant. We see that by only using Padé approximants corresponding to well-conditioned Hankel systems, a weakly stable algorithm is obtained.

4. Algorithm NPADE. Once again, we assume that $B(z) \equiv -1$.[8] We also assume that the norm of $A(z) \pmod{z^{2N}}$ is 1, since we only need the first $2N$ terms

[8] This is no restriction in a theoretical sense, as the general problem can be converted to this form by multiplying both power series by $-B^{-1}(z)$; that this inverse exists follows from the assumption that $B(0) \neq 0$.

of $A(z)$ to calculate the $(N-1, N)$ Padé approximant. This normalization simplifies the presentation of the analysis. With these assumptions, we are now ready to state the algorithm.

ALGORITHM NPADE.

Input: $A(z)$, N, ε, where
N is a positive integer, $N \geq 1$,
$A(z)$ is a real power series with $\| A \| = 1$ (only $A \pmod{z^{2N}}$ is needed), and
ε is a real value, $0 \leq \varepsilon < 1$.

Output: The $(n_k - 1, n_k)$ normalized Padé approximant for $A(z)$, where $n_k \leq N$ is the largest n_k for which the normalized Padé approximant can be computed stably. If $n_k < N$, then an $(N-1, N)$ normalized Padé approximant is also output.

Program:

```
/* Initialization */
$l \leftarrow 0$;
$n_l \leftarrow 0$;
$s_l \leftarrow 1$;
$\mathcal{P}^{(l)} \leftarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$;
/* Compute the Padé Fraction */
While $(n_l + s_l \leq N)$ do
    Compute $\mathcal{R}^{(l)}$ such that $(\mathcal{P}^{(l)}\mathcal{A})\mathrm{mod}(z^{2n_l+2s_l}) = z^{2n_l}\mathcal{R}^{(l)}$;
    Assemble the $2s_l \times 2s_l$ Sylvester matrix $\hat{S}_l$;
    Solve the Sylvester system, using Gaussian elimination with partial pivoting,
        to obtain $\hat{\mathcal{P}}^{(l)}$;
    $\mathcal{P}^{(l+1)} \leftarrow \hat{\mathcal{P}}^{(l)}\mathcal{P}^{(l)}$;
    Rescale $\mathcal{P}^{(l+1)}$ and $\hat{\mathcal{P}}^{(l)}$ so that
        $\| P^{(l+1)} \| + \| Q^{(l+1)} \| = 1$ and $\| U^{(l+1)} \| + \| V^{(l+1)} \| = 1$;
    Compute $r_0^{(l+1)}$ such that $(Q^{(l+1)}A - P^{(l+1)})\mathrm{mod}(z^{2n_l+2s_l+1}) = z^{2n_l+2s_l} r_0^{(l+1)}$;
    If $|r_0^{(l+1)} v_0^{(l+1)}| > \varepsilon$ then
        $n_{l+1} \leftarrow n_l + s_l$;
        $l \leftarrow l + 1$;
        $s_l \leftarrow 1$;
    else
        $s_l \leftarrow s_l + 1$;
End while;
Output $(U^{(l)}, V^{(l)})$;
/* Check the Order of $(U^{(l)}, V^{(l)})$ */
If $(n_l < N)$ then
    /* Output an $(N-1, N)$ Padé approximant */
    $n_{l+1} \leftarrow N$;
    Output $(U^{(l+1)}, V^{(l+1)})$;
```

The algorithm successively computes Padé approximants of type $(n_l - 1, n_l)$, $l = 1, 2, 3, \ldots$. The goal is to select n_l so that the computations are performed accurately. This, as we show in the following sections, is accomplished by means of the stability parameter ε. Assuming that S_{n_l} is a well-conditioned matrix, at the lth step increasing values of s_l are tried until one is found such that, for the Padé approximant pair

$\mathcal{P}^{(l+1)}$, $|r_0^{(l+1)}v_0^{(l+1)}| > \varepsilon$. We see shortly that $r_0^{(l+1)}v_0^{(l+1)}$ is inversely related to the condition number of $S_{n_{l+1}}$, where $n_{l+1} = n_l + s_l$. Indeed, if $r_0^{(l+1)}v_0^{(l+1)} = 0$, then $\hat{S}_{s_l}$ and $S_{n_{l+1}}$ are numerically singular (cf. Theorems 3.2 and 3.4). Thus, the stability parameter ε insists on a certain level of stability of the Sylvester system at the node $(n_{l+1} - 1, n_{l+1})$ before the solution there is accepted. For nodes $(n_l + s - 1, n_l + s)$, $0 < s < s_l$, in between $(n_l - 1, n_l)$ and $(n_{l+1} - 1, n_{l+1})$ the condition numbers of S_{n_l+s}, $0 < s < s_l$, are too large and the solutions there are rejected. We refer to this set of rejected nodes as being in an unstable block of the Padé table (in analogy with a singular block in the c-table (Gragg [31])).

During step l, the algorithm requires the solution of a sequence of Sylvester systems; this is done using Gaussian elimination. If the Sylvester system is singular, the solver is assumed to generate some solution for which $\hat{V}^{(l)}$ and $\hat{Q}^{(l)}$ are nonzero. In this case, $|r_0^{(l+1)}v_0^{(l+1)}|$ is zero (Theorem 3.2) and the solution is rejected. Increasing values of s_l are successively tried until one is found that results in an $|r_0^{(l+1)}v_0^{(l+1)}|$ which is greater than ε. We note that this process succeeds in giving a weakly stable algorithm, but it becomes costly if the stepsize s_l, for any l, becomes large. By taking a larger stepsize, we are obtaining greater numerical stability by sacrificing some efficiency. This tradeoff occurs because, in effect, the algorithm is allowing the loss of some of the regular structure of the system by performing limited partial pivoting in $S_{n_{l+1}}$ (pivoting is performed only within unstable blocks of the Padé table). It remains an open question to find a mechanism which more efficiently determines the smallest s_l for which $S_{n_l+s_l}$ is well conditioned using only information contained in S_{n_l} and $\hat{S}_{s_l}$.

If the While loop exits with $n_l = N$, then we show that the normalized Padé approximant will have a small residual and be close to the true normalized Padé approximant. If the While loop exits with $n_l < N$, then an $(N-1, N)$ normalized Padé approximant is computed outside the loop anyway. If the final system that is solved is nonsingular, then the computed normalized Padé approximant has a small residual, but we make no claims that it is close to the exact normalized Padé approximant. If the last system that is solved is singular, then no claims are made about the computed normalized Padé approximant or its residual.

By assuming exact arithmetic and using $\varepsilon = 0$ for the stability parameter, Algorithm NPADE is equivalent to the method of Cabay and Choi in which the algebraically singular subsystems are skipped over. In this case, $\hat{P}^{(k)}$ and $\hat{Q}^{(k)}$ are simply monomials of degree s and $s+1$, respectively, and $\hat{U}^{(k)}$ and $\hat{V}^{(k)}$ are obtained from the polynomial division of $R^{(k)}$ by $W^{(k)}$. This also analogously describes the algorithms of Rissanen [50], Berlekamp [8], and Massey [45].

5. Summary of results.

The principal result of this paper is the following.

THEOREM 5.1. *Algorithm NPADE is weakly stable over the general class of power series.*

To prove this, we need to show that for a well-conditioned problem (i.e., for a problem that has a corresponding Hankel matrix that is well conditioned), the computed solution (U, V) is close to the exact solution (U_E, V_E). To do this, we derive a bound for the error introduced by one iteration of the algorithm and then apply it iteratively to obtain an expression describing the total error in the computed solution. A forward analysis type of proof is used. Often the problem with this sort of analysis is that the bounds can grow very rapidly (Wilkinson [61]). This is not the case here. Certain relationships that are derived from the order condition (1) allow us to obtain a constant bound on multiplicative terms that would otherwise have led to an exponential growth in the error bounds.

Define

$$(\delta P^{(k)}, \delta Q^{(k)}) = (P^{(k)}, Q^{(k)}) - (P^{(k)}{}_E, Q^{(k)}{}_E) \tag{8}$$

and

$$(\delta U^{(k)}, \delta V^{(k)}) = (U^{(k)}, V^{(k)}) - (U^{(k)}{}_E, V^{(k)}{}_E). \tag{9}$$

Now, rather than satisfying the order condition (1) exactly, instead $\mathcal{P}^{(k)}$ satisfies

$$\mathcal{P}^{(k)}\mathcal{A} = \mathcal{D}^{(k)} + O(z^{2n_k}), \tag{10}$$

where

$$\mathcal{D}^{(k)} = \begin{pmatrix} \delta R^{(k)} \\ \delta W^{(k)} \end{pmatrix}.$$

In $\mathcal{D}^{(k)}$, $\delta R^{(k)}$ and $\delta W^{(k)}$ are polynomials of degree $2n_k - 1$, where the constant and linear terms of $\delta R^{(k)}$ are zero. They describe the error in the order condition for the currently computed Padé approximant pair. With these definitions, then the error bound obtained is the following.

THEOREM 5.2. *If*

$$22(n_l+1)^2(\| \delta R^{(l)} \| + \| \delta W^{(l)} \| + 5(n_l+1)\mu) \le |r_0^{(l)} v_0^{(l)}| \tag{11}$$

holds for $0 \le l \le k$, *then the approximants computed by Algorithm NPADE satisfy*

$$\begin{aligned} \begin{pmatrix} \| \delta P^{(k)} \| + \| \delta Q^{(k)} \| \\ \| \delta U^{(k)} \| + \| \delta V^{(k)} \| \end{pmatrix} &\le \frac{2.2(n_k+1)^3}{\varepsilon^3} \left(\Gamma_0^{(k-1)} + \sum_{l=0}^{k-2} \Gamma_1^{(l)} \right) \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &\quad + O(\mu^2) \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \end{aligned} \tag{12}$$

where

$$\begin{aligned} \Gamma_0^{(l)} &= 4.04(s_l+1)(n_l+1)^3 + 512 s_l^4 \rho_l (n_l+1)^2 \\ &\quad + 8.08(n_{l+1}+1)(n_l+1)(s_l+1)^2, \end{aligned} \tag{13}$$

$$\begin{aligned} \Gamma_1^{(l)} &= 32.32 s_l (s_l+1)(n_l+1)^3(n_{l+1}+1) + 4096 s_l^5 \rho_l (n_l+1)^2 (n_{l+1}+1) \\ &\quad + 32.32(n_{l+1}+1)^3(s_l+1)^2(n_l+1), \end{aligned} \tag{14}$$

ε *is the stability parameter used by Algorithm NPADE, and* ρ_l *is the growth factor associated with the Gaussian elimination method performed at the* l*th step.*

Condition (11) in Theorem 5.2 is a requirement that the error in the order condition not be large. In the derivation of this result, it will be seen that if (11) is satisfied at the lth step and s_l is chosen so that $|r_0^{(l+1)} v_0^{(l+1)}| > \varepsilon$, then the error in the condition will not grow by much at the $(l+1)$st step. Since initially, $\| \delta R^{(0)} \| + \| \delta W^{(0)} \| = 0$, this bound (11) is satisfied in practice.

The Γ_l have large constants and powers of n_l and s_l in them. We believe that these large values are a consequence of the method of analysis and are not a feature of the actual problem being solved.[9] Thus, the bound is crude, but we believe that it still reflects the behavior of the error. Experimental results, summarized later, support this contention. It is also felt that one of the ε's is due to the analysis and is not a feature of the problem itself. This is discussed later. As Wilkinson points out [61, p. 567], "The main object of such an analysis is to expose the potential instabilities, if any, of an algorithm so that hopefully from the insight thus obtained one might be led to improved algorithms. Usually the bound itself is weaker than it might have been because of the necessity of restricting the mass of detail to a reasonable level and because of the limitations imposed by expressing the errors in terms of matrix norms."

What we see in this theorem is that the error grows by a constant amount at each step, so long as $|r_0^{(l)} v_0^{(l)}|$ are chosen so that they are large. Moreover, due to the normalization used, the left-hand side in (12) actually describes the relative error. We will see shortly that $|r_0^{(k)} v_0^{(k)}|$ will be large if the associated Hankel system is well conditioned. Hence, Theorem 5.1 is a consequence of Theorem 5.2.

We now present the analysis that leads to these results. In the next section, we derive bounds for the error $\delta R^{(k)}$ and $\delta W^{(k)}$ in the order condition. In the section following this, we combine these bounds on the error in the order condition with an expression for the inverse of the associated Hankel matrix to obtain bounds on the relative error in the computed solutions.

6. Bounds on the error in the order condition. We first consider what happens during one iteration of Algorithm NPADE. At the kth iteration, the $(n_k - 1, n_k)$ Padé approximant pair $\mathcal{P}^{(k)}$ is available and the algorithm proceeds to compute $\mathcal{P}^{(k+1)}$. For the rest of this paper, the quantities $\mathcal{P}^{(k)}$, $\hat{\mathcal{P}}^{(k)}$, $\mathcal{R}^{(k)}$, etc., are now taken to be the computed quantities that result from the use of floating point operations in Algorithm NPADE. Where an exact quantity is used, it shall be explicitly indicated.

We consider the problem of obtaining bounds for the residual error $\mathcal{D}^{(k)}$ defined in (10). Since we never need to calculate $\mathcal{D}^{(k)}$ explicitly, we actually use the exact relation

$$\mathcal{D}^{(k)} = (\mathcal{P}^{(k)}\mathcal{A}) \pmod{z^{2n_k}}.$$

An iterative step in the algorithm consists of three parts. First, the first $2s_k$ terms of $\mathcal{R}^{(k)}$ are computed. This yields an $\mathcal{R}^{(k)}$ that satisfies

$$\mathcal{P}^{(k)}\mathcal{A} \pmod{z^{2n_k+2s_k}} = \mathcal{D}^{(k)} + z^{2n_k}(\mathcal{R}^{(k)} - \mathcal{C}_1^{(k)}), \tag{15}$$

where $\mathcal{C}_1^{(k)}$ is the error in the computed residual introduced by the floating point operations. Note that in (15), $\mathcal{R}^{(k)}$ and $\mathcal{C}_1^{(k)}$ are polynomials of degree $2s_k - 1$.

[9] Most of these are due to the use of the infinity norm and the inequality given by Lemma 2.6 (e.g., if U and V are polynomials of degree n, then $\| UV \| \leq (n+1)\| U \| \| V \|$). While this inequality can be strict, in practice it is not usually so.

Second, $W^{(k)}$ and $R^{(k)}$ are used to construct a Sylvester matrix

$$(16)\qquad \hat{S}_k = \begin{pmatrix} 0 & 0 & 0 & r_0^{(k)} & 0 & 0 \\ w_0^{(k)} & & 0 & \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots & \vdots & & r_0^{(k)} \\ \vdots & & w_0^{(k)} & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ w_{2s_k-2}^{(k)} & \cdots & w_{s_k-1}^{(k)} & r_{2s_k-1}^{(k)} & \cdots & r_{s_k}^{(k)} \end{pmatrix};$$

an associated system is then solved by Gaussian elimination with partial pivoting to obtain $\hat{\mathcal{P}}^{(k)}$. If we let

$$\mathcal{V} = \begin{pmatrix} \hat{v}_1 \\ \vdots \\ \hat{v}_{s_k} \\ \hat{u}_0 \\ \vdots \\ \hat{u}_{s_k-1} \end{pmatrix}, \quad \mathcal{W} = \begin{pmatrix} w_0 \\ \vdots \\ w_{2s_k-1} \end{pmatrix}, \quad \mathcal{Q} = \begin{pmatrix} \hat{q}_2 \\ \vdots \\ \hat{q}_{s_k+1} \\ 0 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_{s_k} \end{pmatrix},$$

where the components of the above vectors are the coefficients of $\hat{P}^{(k)}$, $\hat{Q}^{(k)}$, $\hat{U}^{(k)}$, $\hat{V}^{(k)}$, and $W^{(k)}(z)$, then $\hat{\mathcal{P}}^{(k)}$ is obtained by solving

$$(17)\qquad \hat{S}_k \mathcal{Q} = \hat{r}_0 e_{2s_k}$$

and

$$(18)\qquad \hat{S}_k \mathcal{V} = -\hat{v}_0 \mathcal{W},$$

where $-\hat{v}_0$ and $\hat{r}_0$ are parameters (keeping in mind that $\hat{p}_0 = \hat{p}_1 = \hat{q}_0 = \hat{q}_1 = 0$).[10] If $\hat{S}_k$ is nonsingular, then $-\hat{v}_0$ and $\hat{r}_0$ are set to 1 and (17) and (18) are solved. If $\hat{S}_k$ is singular, then a nontrivial solution is obtained for which $-\hat{v}_0$ and $\hat{r}_0$ are 0. (The solution that results from a singular system yields $|r_0^{k+1} v_0^{k+1}| = 0$, and so is only used if this is the last system solved by the algorithm.) This computed $\hat{\mathcal{P}}^{(k)}$ is not the exact solution of the equivalent problem $\hat{\mathcal{P}}^{(k)}\mathcal{R}^{(k)} = O(z^{2s_k})$. Rather, $\hat{\mathcal{P}}^{(k)}$ satisfies

$$(19)\qquad \hat{\mathcal{P}}^{(k)}\mathcal{R}^{(k)} = \mathcal{C}_2^{(k)} + O(z^{2s_k}).$$

Third, $\mathcal{P}^{(k+1)}$ is obtained from $fl(\hat{\mathcal{P}}^{(k)}\mathcal{P}^{(k)})$, or, equivalently,

$$(20)\qquad \mathcal{P}^{(k+1)} = \hat{\mathcal{P}}^{(k)}\mathcal{P}^{(k)} + \mathcal{C}_3^{(k)},$$

where $\mathcal{C}_3^{(k)}$ is the error introduced by the floating point operations. At this point, $\mathcal{P}^{(k+1)}$ and $\hat{\mathcal{P}}^{(k)}$ are scaled so that

$$\| P^{(k+1)} \| + \| Q^{(k+1)} \| = 1$$

[10] The case $s_k = 1$ is covered as well, as then $\hat{P}^{(k)} = 0$ and $\hat{Q}^{(k)} = z^2/w_0$.

and

$$\| U^{(k+1)} \| + \| V^{(k+1)} \| = 1.$$

We assume that this introduces no additional error. This is a reasonable assumption because in our analysis, we only actually require 1 to be an upper bound for the above quantities. We could modify the normalizations and rescaling so that they could be done exactly. However, since we are dealing with the infinity norm and the errors that result from this normalization are small, we have made this assumption for the sake of simplicity. Now, this scaling will have no effect on the error analysis as $\hat{r}_0$ and $\hat{v}_0$ are just parameters in the solution of (17) and (18). In other words, we could resolve the system using the new values of $\hat{r}_0$ and $\hat{v}_0$ instead of 1; the error bounds will be the same.

A straightforward error analysis of these three parts produces bounds on the $\mathcal{C}_j^{(k)}$ in (15), (19), and (20). As a notational convenience, we use $| \cdot |$ to denote the componentwise application of the infinity norm to vectors and matrices. For example,

$$|\hat{\mathcal{P}}^{(k)}| = \begin{pmatrix} \| \hat{P}^{(k)} \| & \| \hat{Q}^{(k)} \| \\ \| \hat{U}^{(k)} \| & \| \hat{V}^{(k)} \| \end{pmatrix}.$$

We also rely on certain normalizations that have been applied to various quantities, namely,

$$\| P^{(k)} \| + \| Q^{(k)} \| = 1, \tag{21}$$

$$\| U^{(k)} \| + \| V^{(k)} \| = 1, \tag{22}$$

$$\| A \| \pmod{z^{2N}} = 1. \tag{23}$$

We state the results for all three parts, but only give the details of the proof for one lemma. The remaining proofs are similar and can be found in Meleshko [46], [47].

LEMMA 6.1. *The computed $\mathcal{R}^{(k)}$ satisfies*

$$z^{2n_k}\mathcal{R}^{(k)} = \mathcal{P}^{(k)}\mathcal{A} \pmod{z^{2n_k+2s_k}} - \mathcal{D}^{(k)} + z^{2n_k}\mathcal{C}_1^{(k)} \tag{24}$$

where

$$|\mathcal{C}_1^{(k)}| \leq 1.01(n_k+1)^2\mu \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{25}$$

Proof. This follows from the fact that

$$z^{2n_k}\mathcal{R}^{(k)} = \begin{pmatrix} fl((AQ^{(k)}) & (\text{mod } z^{2n_k+2s_k})) \\ fl((AV^{(k)}) & (\text{mod } z^{2n_k+2s_k})) \end{pmatrix} - \begin{pmatrix} fl((AQ^{(k)}) & (\text{mod } z^{2n_k})) \\ fl((AV^{(k)}) & (\text{mod } z^{2n_k})) \end{pmatrix},$$

using (21), (22), (23), and Lemma 2.7. □

LEMMA 6.2. *If $\hat{S}_k$ is nonsingular*[11] *and $\hat{\mathcal{P}}^{(k)}$ is obtained by solving* (17) *and* (18) *using Gaussian elimination with partial pivoting, then*

$$\hat{\mathcal{P}}^{(k)}\mathcal{R}^{(k)} = \mathcal{C}_2^{(k)} + O(z^{2s_k}), \tag{26}$$

[11] We believe that this result is true even if $\hat{S}_k$ is singular as the singular case can be handled by setting $\hat{r}_0 = \hat{v}_0 = 0$ and then solving an underdetermined homogeneous system of equations.

where

$$(27) \qquad |\mathcal{C}_2^{(k)}| \leq (128 s_k^4 \rho_k (n_k+1)\mu + O(\mu^2)) \begin{pmatrix} \max\{\| \hat{P}^{(k)} \|, \| \hat{Q}^{(k)} \|\} \\ \max\{\| \hat{U}^{(k)} \|, \| \hat{V}^{(k)} \|\} \end{pmatrix}$$

and ρ_k is a constant of order unity in practice.

Proof. When (17) and (18) are solved using Gaussian elimination with partial pivoting, the solutions obtained are exact solutions of

$$(28) \qquad (\hat{S}_k + E_Q)\hat{Q} = \hat{r}_0 e_{2s_k}$$

and

$$(29) \qquad (\hat{S}_k + E_V)\hat{V} = -\hat{v}_0 W,$$

where

$$\| E_{(\cdot)} \| \leq 8(2s_k)^3 \rho_k \| \hat{S}_k \| \mu + O(\mu^2)$$

and ρ_k is the growth factor associated with the LU decomposition of $\hat{S}_k$ (Golub and Van Loan [30, p. 67]). With partial pivoting, ρ_k is of order unity in practice.[12] Now from (28) and (29), we have

$$(30) \qquad \hat{S}_k\hat{Q} - \hat{r}_0 e_{2s_k} = -E_Q\hat{Q}$$

and

$$(31) \qquad \hat{S}_k\hat{V} + \hat{v}_0 W = -E_V\hat{V},$$

where

$$\begin{pmatrix} \| E_Q\hat{Q} \| \\ \| E_V\hat{V} \| \end{pmatrix} \leq (64 s_k^3 \rho_k \| \hat{S}_k \| \mu + O(\mu^2)) \begin{pmatrix} \max\{\| \hat{P}^{(k)} \|, \| \hat{Q}^{(k)} \|\} \\ \max\{\| \hat{U}^{(k)} \|, \| \hat{V}^{(k)} \|\} \end{pmatrix}.$$

It is easily shown, from Lemmas 2.6 and 6.1, that

$$\| \hat{S}_k \| \leq 2s_k(n_k+1)(1+1.01(n_k+1)\mu).$$

Using this and absorbing the terms that are nonlinear in μ into the $O(\mu^2)$ term, the expression on the right-hand side of (27) is obtained. Now, since a $\hat{\mathcal{P}}^{(k)}$ which satisfies $\hat{\mathcal{P}}^{(k)}\mathcal{R}^{(k)} = O(z^{2s_k})$ can be found by solving (17) and (18) exactly, it is easy to see the correspondence between (26) and the linear systems given by (28) and (29), and the result follows. □

LEMMA 6.3. *If $\mathcal{P}^{(k+1)} = fl(\hat{\mathcal{P}}^{(k)}\mathcal{P}^{(k)})$, then*

$$(32) \qquad \mathcal{P}^{(k+1)} = \hat{\mathcal{P}}^{(k)}\mathcal{P}^{(k)} + \mathcal{C}_3^{(k)},$$

where

$$(33) \qquad |\mathcal{C}_3^{(k)}| \leq (2.02(s_k+1)^2\mu + O(\mu^2))|\hat{\mathcal{P}}^{(k)}||\mathcal{P}^{(k)}|.$$

[12] Examples can be constructed where the growth factor ρ grows exponentially if partial pivoting is used, but in practice ρ is usually comparable to the modest growth that results when complete pivoting is used (which is approximately 10 in practice)(Golub and Van Loan [30, p. 69]). Further discussion and new results regarding the growth factor ρ can be found in Higham and Higham [34] and Trefethen and Schreiber [54].

Proof. This follows from Lemmas 2.7 and 2.8 and the definition of $\mathcal{P}^{(k+1)}$. □

Use of the above expressions enables us to express $\mathcal{D}^{(k+1)}$ in terms of $\mathcal{D}^{(k)}$ and the errors introduced in one iteration.

LEMMA 6.4. *Let the residual error $\mathcal{D}^{(k)}$ in the solution $\mathcal{P}^{(k)}$ at the start of the kth iteration satisfy*

$$\mathcal{P}^{(k)}\mathcal{A} = \mathcal{D}^{(k)} + O(z^{2n_k}). \tag{34}$$

Then the residual error $\mathcal{D}^{(k+1)}$ after k complete iterations satisfies

$$\mathcal{D}^{(k+1)} = \hat{\mathcal{P}}^{(k)}\mathcal{D}^{(k)} + \mathcal{L}^{(k)}, \tag{35}$$

where

$$\mathcal{L}^{(k)} = \mathcal{L}_1^{(k)} + \mathcal{L}_2^{(k)} + \mathcal{L}_3^{(k)}$$

and

$$\mathcal{L}_1^{(k)} = -z^{2n_k}\hat{\mathcal{P}}^{(k)}\mathcal{C}_1^{(k)} \pmod{z^{2n_{k+1}}}, \tag{36}$$

$$\mathcal{L}_2^{(k)} = z^{2n_k}\mathcal{C}_2^{(k)}, \text{ and} \tag{37}$$

$$\mathcal{L}_3^{(k)} = \mathcal{C}_3^{(k)}\mathcal{A} \pmod{z^{2n_{k+1}}}. \tag{38}$$

Proof. We have using, in order, (32), (38), (24), (36), (26), and (37),

$$\begin{aligned}
\mathcal{D}^{(k+1)} &= (\mathcal{P}^{(k+1)}\mathcal{A}) \pmod{z^{2n_{k+1}}}\\
&= ((\hat{\mathcal{P}}^{(k)}\mathcal{P}^{(k)} + \mathcal{C}_3^{(k)})\mathcal{A}) \pmod{z^{2n_{k+1}}}\\
&= (\hat{\mathcal{P}}^{(k)}(\mathcal{P}^{(k)}\mathcal{A})) \pmod{z^{2n_{k+1}}} + \mathcal{L}_3^{(k)}\\
&= (\hat{\mathcal{P}}^{(k)}(z^{2n_k}\mathcal{R}^{(k)} + \mathcal{D}^{(k)} - z^{2n_k}\mathcal{C}_1^{(k)})) \pmod{z^{2n_{k+1}}} + \mathcal{L}_3^{(k)}\\
&= (z^{2n_k}(\hat{\mathcal{P}}^{(k)}\mathcal{R}^{(k)}) + \hat{\mathcal{P}}^{(k)}\mathcal{D}^{(k)}) \pmod{z^{2n_{k+1}}} + \mathcal{L}_1^{(k)} + \mathcal{L}_3^{(k)}\\
&= (z^{2n_k}(\mathcal{C}_2^{(k)} + O(z^{2s_k})) \pmod{z^{2n_{k+1}}}) + \hat{\mathcal{P}}^{(k)}\mathcal{D}^{(k)} + \mathcal{L}_1^{(k)} + \mathcal{L}_3^{(k)}\\
&= \hat{\mathcal{P}}^{(k)}\mathcal{D}^{(k)} + \mathcal{L}_1^{(k)} + \mathcal{L}_2^{(k)} + \mathcal{L}_3^{(k)}. \qquad \Box
\end{aligned}$$

Thus the residual $\mathcal{D}^{(k+1)}$ is composed of the error $\mathcal{L}^{(k)}$ introduced by the kth iteration plus the residual error $\mathcal{D}^{(k)}$ from the previous step propagated by $\hat{\mathcal{P}}^{(k)}$. Now if the relationship (35) is applied recursively, we get the following.

THEOREM 6.5.

$$\mathcal{D}^{(k+1)} = \sum_{l=1}^{k+1} \mathcal{G}_l^{(k)}\mathcal{L}^{(l-1)}, \qquad k \geq 0, \tag{39}$$

where

$$\begin{aligned}
\mathcal{G}_l^{(k)} &= \hat{\mathcal{P}}^{(k)}\hat{\mathcal{P}}^{(k-1)}\cdots\hat{\mathcal{P}}^{(l)}, \qquad 0 \leq l \leq k, \text{ and}\\
\mathcal{G}_{k+1}^{(k)} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.
\end{aligned}$$

Proof. This follows from Lemma 6.4 by induction. □

From (39), we see that $\mathcal{G}_l^{(k)}$ describes the way that an error introduced at the $(l-1)$th iteration in the algorithm is propagated through to the kth iteration.

We can also take (39) and rewrite it so that the errors associated with each part of an iteration are given separately. This gives us the following.

COROLLARY 6.6.

$$\mathcal{D}^{(k+1)} = \mathcal{E}_1^{(k)} + \mathcal{E}_2^{(k)} + \mathcal{E}_3^{(k)}, \tag{40}$$

where

$$\mathcal{E}_1^{(k)} = \sum_{l=1}^{k} -z^{2n_l}\mathcal{G}_{l+1}^{(k)}(\hat{\mathcal{P}}^{(l)}\mathcal{C}_1^{(l)} \pmod{z^{2s_l}}), \tag{41}$$

$$\mathcal{E}_2^{(k)} = \sum_{l=0}^{k} z^{2n_l}\mathcal{G}_{l+1}^{(k)}\mathcal{C}_2^{(l)}, \textit{ and} \tag{42}$$

$$\mathcal{E}_3^{(k)} = \sum_{l=1}^{k} \mathcal{G}_{l+1}^{(k)}((\mathcal{C}_3^{(l)}\mathcal{A}) \pmod{z^{2n_{l+1}}}). \tag{43}$$

In (41) and (43), we have used some results that are implicit from the statement of Algorithm NPADE, namely, $\mathcal{C}_1^{(0)} = \mathcal{C}_3^{(0)} = 0$. To see this, we briefly consider the first pass through the While loop in the algorithm. In the first pass,

$$\mathcal{P}^{(0)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

so that $\mathcal{R}^{(0)} = \mathcal{A} \pmod{z^{2s_0}}$ exactly and there is no error in computing this first residual; i.e., $\mathcal{C}_1^{(0)} = 0$. Once the algorithm has computed $\hat{\mathcal{P}}^{(0)}$, it then computes $fl(\hat{\mathcal{P}}^{(0)}\mathcal{P}^{(0)})$. Again, since $\mathcal{P}^{(0)}$ is the identity matrix, we have no error in this step, so $\mathcal{C}_3^{(0)} = 0$. Also, if $s_0 = n$, then the error is given by (42) above, which is just the error that results from Gaussian elimination.

Returning to our primary goal of obtaining a bound for the total error, we observe that the key expression to bound is $\mathcal{G}_l^{(k)}$. A concern here is that the $\hat{\mathcal{P}}^{(j)}$'s making up $\mathcal{G}_l^{(k)}$ will cause exponential growth with k in $\mathcal{G}_l^{(k)}$. However, this is not the case. It will be seen that the bound on $\mathcal{G}_l^{(k)}$ is independent of k. Hence, the error as described by (39) will be the sum of the errors introduced in each iteration l. Thus, in this sense, the error grows additively; that is, the error $\mathcal{L}^{(l)}$ at iteration l that is propagated to iteration k by $\mathcal{G}_{l+1}^{(k)}$ does not grow with k.

To obtain this bound on $\mathcal{G}_i^{(k)}$, we first observe that (10) can be rewritten as

$$AQ^{(k)} - P^{(k)} = \delta R^{(k)} + z^{2n_k}R^{(k)} \tag{44}$$

and

$$AV^{(k)} - U^{(k)} = \delta W^{(k)} + z^{2n_k}W^{(k)}, \tag{45}$$

where $v_0^{(k)} \neq 0$ and $r_0^{(k)} \neq 0$. Multiplying (45) by $Q^{(k)}$ and (44) by $V^{(k)}$ and then subtracting (45) from (44) gives us

$$Q^{(k)}U^{(k)} - V^{(k)}P^{(k)} = V^{(k)}\delta R^{(k)} - Q^{(k)}\delta W^{(k)} + z^{2n_k}(V^{(k)}R^{(k)} - Q^{(k)}W^{(k)}). \tag{46}$$

Now the degree of $(Q^{(k)}U^{(k)} - V^{(k)}P^{(k)})$ is less than or equal to $2n_k$ so all higher-order terms on the right-hand side of (46) must cancel. Since $q_0^{(k)} = 0$, what we have is

$$Q^{(k)}U^{(k)} - V^{(k)}P^{(k)} = \vartheta^{(k)}(z) + v_0 r_0 z^{2n_k}, \tag{47}$$

where

$$\vartheta^{(k)}(z) = (V^{(k)}\delta R^{(k)} - Q^{(k)}\delta W^{(k)}) \bmod (z^{2n_k+1}). \tag{48}$$

Letting

$$\mathcal{P}_I^{(k)} = \begin{pmatrix} V^{(k)} & -Q^{(k)} \\ -U^{(k)} & P^{(k)} \end{pmatrix}$$

and using (47), we obtain

$$\mathcal{P}^{(k)}\mathcal{P}_I^{(k)} = -\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}(r_0^{(k)}v_0^{(k)}z^{2n_k} + \vartheta^{(k)}(z)). \tag{49}$$

With this, we are now able to prove the following lemma.

LEMMA 6.7. *Assuming that $r_0^{(k)}$ and $v_0^{(k)}$ are large and $\| \delta R^{(k)} \|$ and $\| \delta W^{(k)} \|$ are small so that*

$$22(n_k+1)^2(\| \delta R^{(k)} \| + \| \delta W^{(k)} \| + 5(n_k+1)\mu) \le |r_0^{(k)}v_0^{(k)}|, \tag{50}$$

then

$$|\mathcal{G}_k^{(k)}|\begin{pmatrix} 1 \\ 1 \end{pmatrix} \le \frac{4(n_k+1)}{|r_0^{(k)}v_0^{(k)}|}\begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{51}$$

Proof. From (32), we have that

$$|\mathcal{P}^{(k+1)}\mathcal{P}_I^{(k)}|\begin{pmatrix} 1 \\ 1 \end{pmatrix} = |\hat{\mathcal{P}}^{(k)}\mathcal{P}^{(k)}\mathcal{P}_I^{(k)} + \mathcal{C}_3^{(k)}\mathcal{P}_I^{(k)}|\begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{52}$$

Now, using (21) and (22),

$$|\mathcal{P}^{(k+1)}||\mathcal{P}_I^{(k)}|\begin{pmatrix} 1 \\ 1 \end{pmatrix} \le |\mathcal{P}^{(k+1)}|\begin{pmatrix} 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \tag{53}$$

so that the left-hand side of (52) is bounded by

$$\begin{aligned} |\mathcal{P}^{(k+1)}\mathcal{P}_I^{(k)}|\begin{pmatrix} 1 \\ 1 \end{pmatrix} &\le (n_k+1)|\mathcal{P}^{(k+1)}||\mathcal{P}_I^{(k)}|\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &\le 2(n_k+1)\begin{pmatrix} 1 \\ 1 \end{pmatrix}. \end{aligned} \tag{54}$$

On the other hand, using (49), Lemma 2.7, and (33), the right-hand side of (52) is bounded below by

$$|\hat{\mathcal{P}}^{(k)}\mathcal{P}^{(k)}\mathcal{P}_I^{(k)} + \mathcal{C}_3^{(k)}\mathcal{P}_I^{(k)}|\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\begin{aligned}
&\geq |\hat{\mathcal{P}}^{(k)}(|r_0^{(k)}v_0^{(k)}|z^{2n_k})|\begin{pmatrix}1\\1\end{pmatrix} - |\hat{\mathcal{P}}^{(k)}\vartheta^{(k)}(z)|\begin{pmatrix}1\\1\end{pmatrix}\\
&\quad - (n_k+1)|\mathcal{C}_3^{(k)}||\mathcal{P}_I^{(k)}|\begin{pmatrix}1\\1\end{pmatrix} \qquad (55)\\
&\geq |r_0^{(k)}v_0^{(k)}||\hat{\mathcal{P}}^{(k)}|\begin{pmatrix}1\\1\end{pmatrix} - (s_k+1)\|\ \vartheta^{(k)}(z)\ \|\cdot|\hat{\mathcal{P}}^{(k)}|\begin{pmatrix}1\\1\end{pmatrix}\\
&\quad - (n_k+1)[2.02(s_k+1)^2\mu + O(\mu^2)]|\hat{\mathcal{P}}^{(k)}||\mathcal{P}^{(k)}||\mathcal{P}_I^{(k)}|\begin{pmatrix}1\\1\end{pmatrix}.
\end{aligned}$$

Now, using the same argument as used in (53), (55) gives us

(56)

$$\begin{aligned}
&|\hat{\mathcal{P}}^{(k)}\mathcal{P}^{(k)}\mathcal{P}_I^{(k)} + \mathcal{C}_3^{(k)}\mathcal{P}_I^{(k)}|\begin{pmatrix}1\\1\end{pmatrix}\\
&\quad\geq |r_0^{(k)}v_0^{(k)}||\hat{\mathcal{P}}^{(k)}|\begin{pmatrix}1\\1\end{pmatrix} - (s_k+1)(n_k+1)(\|\ \delta R^{(k)}\ \| + \|\ \delta W^{(k)}\ \|)|\hat{\mathcal{P}}^{(k)}|\begin{pmatrix}1\\1\end{pmatrix}\\
&\qquad - (4.04(n_k+1)(s_k+1)^2\mu + O(\mu^2))|\hat{\mathcal{P}}^{(k)}|\begin{pmatrix}1\\1\end{pmatrix}\\
&\quad= \Big(|r_0^{(k)}v_0^{(k)}| - [(s_k+1)(n_k+1)(\|\ \delta R^{(k)}\ \| + \|\ \delta W^{(k)}\ \|\\
&\qquad + 4.04(s_k+1)\mu + O(\mu^2))]\Big)|\hat{\mathcal{P}}^{(k)}|\begin{pmatrix}1\\1\end{pmatrix}\\
&\quad\geq |\hat{\mathcal{P}}^{(k)}|\begin{pmatrix}1\\1\end{pmatrix}\frac{|r_0^{(k)}v_0^{(k)}|}{2}.
\end{aligned}$$

The last inequality follows from the application of (50), if we assume that $s_k \leq n_k$. But the above argument also holds, with minor modifications, if $s_k > n_k$; for example, the first occurrence of (s_k+1) in (55) can be replaced by $(2n_k+1)$ and the second by (n_k+1). Thus (56) is valid in general.[13]

Since $\mathcal{G}_k^{(k)} = \hat{\mathcal{P}}^{(k)}$, it follows from (54), (56), and (52), that

$$|\mathcal{G}_k^{(k)}|\begin{pmatrix}1\\1\end{pmatrix} \leq \frac{4(n_k+1)}{|r_0^{(k)}v_0^{(k)}|}\begin{pmatrix}1\\1\end{pmatrix}. \qquad \square$$

To simplify our notation, let

$$\sigma^{(k)} = r_0^{(k)}v_0^{(k)}. \tag{57}$$

COROLLARY 6.8. *Given that* (50) *holds,*

$$|\mathcal{C}_3^{(k)}|\begin{pmatrix}1\\1\end{pmatrix} \leq \left(\frac{8.08(s_k+1)^2(n_k+1)}{|\sigma^{(k)}|}\mu + O(\mu^2)\right)\begin{pmatrix}1\\1\end{pmatrix}. \tag{58}$$

Proof. This result is immediate from (33) and Lemma 6.7. □

These two results now enable us to obtain a bound for the general $\mathcal{G}_l^{(k)}$. This bound essentially says that $|\mathcal{G}_l^{(k)}|$ is independent of k. In other words, when the error

[13] The application of (50) only requires a constant factor of 2 (or 4 if $s_k > n_k$) instead of 22. However, the larger factor is required later in Lemma 7.5.

is introduced, it does not grow significantly after its introduction. A consequence of this is that the errors that occur in each step are additive.

THEOREM 6.9. *If the conditions of Lemma* 6.7 *hold for each* $l \le k$, *then*

$$|\mathcal{G}_l^{(k)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \le \frac{4(n_l+1)+O(\mu)}{|\sigma^{(l)}|} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{59}$$

Proof. By repeated application of the recurrence relation (32), we have

$$\mathcal{P}^{(k+1)} = \mathcal{G}_i^{(k)} \mathcal{P}^{(i)} + \sum_{l=i}^{k} \mathcal{G}_{l+1}^{(k)} \mathcal{C}_3^{(l)}. \tag{60}$$

The result is then proved by induction on i, starting with $i = k$. For $i = k$, (59) reduces to (51).

Assume that (59) holds for $\mathcal{G}_k^{(k)}, \mathcal{G}_{k-1}^{(k)}, \ldots, \mathcal{G}_i^{(k)}$. From (60), it follows that

$$|\mathcal{P}^{(k+1)}\mathcal{P}_I^{(i-1)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \left| \mathcal{G}_{i-1}^{(k)} \mathcal{P}^{(i-1)} \mathcal{P}_I^{(i-1)} + \sum_{l=i-1}^{k} \mathcal{G}_{l+1}^{(k)} \mathcal{C}_3^{(l)} \mathcal{P}_I^{(i-1)} \right| \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

As in the proof of Lemma 6.7, we can show that

$$|\mathcal{P}^{(k+1)}\mathcal{P}_I^{(i-1)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \le 2(n_{i-1}+1) \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

On the other hand, using

$$\mathcal{P}^{(i-1)}\mathcal{P}_I^{(i-1)} = -\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (r_0^{(i-1)} v_0^{(i-1)} z^{2n_{i-1}} + \vartheta^{(i-1)}(z)), \tag{61}$$

(21), (22), (58), the fact that the degree of $\vartheta^{(i-1)}$ is bounded by $2n_{i-1}$, and the inductive hypothesis,

$$\begin{aligned}
&\left| \mathcal{G}_{i-1}^{(k)} \mathcal{P}^{(i-1)} \mathcal{P}_I^{(i-1)} + \sum_{l=i-1}^{k} \mathcal{G}_{l+1}^{(k)} \mathcal{C}_3^{(l)} \mathcal{P}_I^{(i-1)} \right| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&\quad \ge |\mathcal{G}_{i-1}^{(k)} \left(-\sigma^{(i-1)} z^{2n_{i-1}} - \vartheta^{(i-1)}(z)\right)| \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \sum_{l=i-1}^{k} |\mathcal{G}_{l+1}^{(k)} \mathcal{C}_3^{(l)} \mathcal{P}_I^{(i-1)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&\quad \ge |\sigma^{(i-1)}||\mathcal{G}_{i-1}^{(k)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} - (2n_{i-1}+1)|\mathcal{G}_{i-1}^{(k)}| \cdot \| \vartheta^{(i-1)}(z) \| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&\qquad - \sum_{l=i-1}^{k} (n_{i-1}+1)|\mathcal{G}_{l+1}^{(k)} \mathcal{C}_3^{(l)}||\mathcal{P}_I^{(i-1)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&\quad \ge |\sigma^{(i-1)}||\mathcal{G}_{i-1}^{(k)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&\qquad - (2n_{i-1}+1)(n_{i-1}+1)(\| \delta R^{(i-1)} \| + \| \delta W^{(i-1)} \|)|\mathcal{G}_{i-1}^{(k)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&\qquad - \sum_{l=i-1}^{k} (n_{i-1}+1)(n_{l+1}+1)|\mathcal{G}_{l+1}^{(k)}||\mathcal{C}_3^{(l)}||\mathcal{P}_I^{(i-1)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix}
\end{aligned}$$

$$\geq |\mathcal{G}_{i-1}^{(k)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} (|\sigma^{(i-1)}| - (2n_{i-1}+1)(n_{i-1}+1)(\| \delta R^{(i-1)} \| + \| \delta W^{(i-1)} \|))$$

$$- \sum_{l=i-1}^{k} 2(n_{i-1}+1)(n_{l+1}+1)$$

$$\cdot \left(\frac{4(n_{l+1}+1)+O(\mu)}{|\sigma^{(l+1)}|} \right) \left(\frac{8.08(s_l+1)^2(n_l+1)\mu + O(\mu^2)}{|\sigma^{(l)}|} \right) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\geq \frac{|\sigma^{(i-1)}|}{2} |\mathcal{G}_{i-1}^{(k)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} - O(\mu) \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

so that

$$|\mathcal{G}_{i-1}^{(k)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \leq \frac{4(n_{i-1}+1)+O(\mu)}{|\sigma^{(i-1)}|} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and the result follows by induction. □

In the above theorem, we have taken the liberty of replacing a summation involving terms linear in μ with an $O(\mu)$ expression. We could have left the summation in explicitly, but, as we shall see, this summation becomes quadratic in μ when it is used to obtain a bound on $\mathcal{D}^{(k)}$.

LEMMA 6.10. *If the assumptions of Theorem* 6.9 *hold, then*

$$|\mathcal{E}_1^{(k)}| \leq \left(\frac{4.04(s_k+1)(n_k+1)^3}{|\sigma^{(k)}|} \right.$$

$$\left. + \sum_{l=0}^{k-1} \frac{32.32 s_l(s_l+1)(n_{l+1}+1)(n_l+1)^3}{|\sigma^{(l)}\sigma^{(l+1)}|} \right) \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{62}$$

$$+ O(\mu^2) \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$|\mathcal{E}_2^{(k)}| \leq \left(\frac{512 s_k^4 \rho_k (n_k+1)^2}{|\sigma^{(k)}|} \right.$$

$$\left. + \sum_{l=0}^{k-1} \frac{4096 s_l^5 \rho_l (n_l+1)^2 (n_{l+1}+1)}{|\sigma^{(l)}\sigma^{(l+1)}|} \right) \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{63}$$

$$+ O(\mu^2) \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$|\mathcal{E}_3^{(k)}| \leq \left(\frac{8.08(n_{k+1}+1)(n_k+1)(s_k+1)^2}{|\sigma^{(k)}|} \right.$$

$$\left. + \sum_{l=0}^{k-1} \frac{32.32(n_{l+1}+1)^3(s_l+1)^2(n_l+1)}{|\sigma^{(l)}\sigma^{(l+1)}|} \right) \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{64}$$

$$+ O(\mu^2) \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Proof. These results follow from the bounds obtained for the $\mathcal{C}_i^{(l)}$ and Theorem 6.9. We derive the bound for $\mathcal{E}_2^{(k)}$ only (defined explicitly in (42)); the other two follow by even simpler arguments. From (42), Lemma 2.7, (19), (27), (59), in that order, and the fact that

$$\begin{pmatrix} \max\{\| \hat{P}^{(i)} \|, \| \hat{Q}^{(i)} \|\} \\ \max\{\| \hat{U}^{(i)} \|, \| \hat{V}^{(i)} \|\} \end{pmatrix} \le |\mathcal{G}_i^{(i)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{65}$$

$$\begin{aligned}
|\mathcal{E}_2^{(k)}| &\le |\mathcal{C}_2^{(k)}| + \sum_{l=0}^{k-1} |\mathcal{G}_{l+1}^{(k)} \mathcal{C}_2^{(l)}| \le |\mathcal{C}_2^{(k)}| + \sum_{l=0}^{k-1} 2s_l |\mathcal{G}_{l+1}^{(k)}||\mathcal{C}_2^{(l)}| \\
&\le (128 s_k^4 \rho_k (n_k+1)\mu + O(\mu^2)) \begin{pmatrix} \max\{\| \hat{P}^{(k)} \|, \| \hat{Q}^{(k)} \|\} \\ \max\{\| \hat{U}^{(k)} \|, \| \hat{V}^{(k)} \|\} \end{pmatrix} \\
&\quad + \sum_{l=0}^{k-1} (256 s_l^5 \rho_l (n_l+1)\mu + O(\mu^2)) |\mathcal{G}_{l+1}^{(k)}| \begin{pmatrix} \max\{\| \hat{P}^{(l)} \|, \| \hat{Q}^{(l)} \|\} \\ \max\{\| \hat{U}^{(l)} \|, \| \hat{V}^{(l)} \|\} \end{pmatrix} \\
&\le (128 s_k^4 \rho_k (n_k+1)\mu + O(\mu^2)) |\mathcal{G}_k^{(k)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&\quad + \sum_{l=0}^{k-1} (256 s_l^5 \rho_l (n_l+1)\mu + O(\mu^2)) |\mathcal{G}_{l+1}^{(k)}||\mathcal{G}_l^{(l)}| \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&\le \frac{512 s_k^4 \rho_k (n_k+1)^2 \mu + O(\mu^2)}{|\sigma^{(k)}|} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&\quad + \sum_{l=0}^{k-1} (256 s_l^5 \rho_l (n_l+1)\mu + O(\mu^2)) \frac{16(n_l+1)(n_{l+1}+1) + O(\mu^2)}{|\sigma^{(l)}\sigma^{(l+1)}|} \begin{pmatrix} 1 \\ 1 \end{pmatrix}
\end{aligned}$$

and the result follows. □

In the above lemma, the bounds obtained involve the products $\sigma^{(l)}\sigma^{(l+1)}$. These result from inequalities involving the expression $|\mathcal{G}_{l+1}^{(k)}||\mathcal{G}_l^{(l)}|$, which in turn resulted from the application of (65). However, it is seen that $\mathcal{G}_{l+1}^{(k)}\mathcal{G}_l^{(l)} = \mathcal{G}_l^{(k)}$, so it is felt that the inequalities are crude and the bounds should just involve a single $\sigma^{(l)}$. Experimental results (Meleshko and Cabay [48]) support this conjecture.

Finally, we have the following theorem.

THEOREM 6.11. *If the conditions of Lemma* 6.7 *hold for each $l \le k$, then*

$$\begin{aligned}
|\mathcal{D}^{(k+1)}| \le &\left(\frac{\Gamma_0^{(k)}}{|\sigma^{(k)}|} + \sum_{l=0}^{k-1} \frac{\Gamma_1^{(l)}}{|\sigma^{(l)}\sigma^{(l+1)}|} \right) \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&+ O(\mu^2) \begin{pmatrix} 1 \\ 1 \end{pmatrix},
\end{aligned} \tag{66}$$

where $\Gamma_0^{(l)}$ and $\Gamma_1^{(l)}$ are as defined in (13) *and* (14).

Proof. Since $\mathcal{D}^{(k+1)} = \mathcal{E}_1^{(k)} + \mathcal{E}_2^{(k)} + \mathcal{E}_3^{(k)}$, application of Lemma 6.10 yields

$$\begin{aligned}
|\mathcal{D}^{(k+1)}| \le &\Big(4.04(s_k+1)(n_k+1)^3 + 512 s_k^4 \rho_k (n_k+1)^2 \\
&+ 8.08(n_{k+1}+1)(n_k+1)(s_k+1)^2\Big) \frac{\mu}{|\sigma^{(k)}|} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&+ \sum_{l=0}^{k-1} \Bigg(\big(32.32 s_l (s_l+1)(n_l+1)^3 (n_{l+1}+1) + 4096 s_l^5 \rho_l (n_l+1)^2 (n_{l+1}+1)
\end{aligned}$$

$$+32.32(n_{l+1}+1)^3(s_l+1)^2(n_l+1))\frac{\mu}{|\sigma^{(l)}\sigma^{(l+1)}|}\Big)\begin{pmatrix}1\\1\end{pmatrix}$$
$$+O(\mu^2)\begin{pmatrix}1\\1\end{pmatrix}$$

and the result follows. □

Observe that each term in the summation in (66) is independent of k. Also, Theorem 6.11 assures us that if $|\mathcal{D}^{(k)}|$ is small and $|r_0^{(k)}|$ and $|v_0^{(k)}|$ are large, then $|\mathcal{D}^{(k+1)}|$ will also be small. Moreover, if s_k is chosen so that $|r_0^{(k+1)}|$ and $|v_0^{(k+1)}|$ are large as well, then Theorem 6.11 can be applied again at the next iteration. Thus, $|\mathcal{D}^{(k)}|$ will remain small for all k, as long as s_k can be chosen in such a fashion. In experiments with power series whose coefficients were randomly generated (Meleshko and Cabay [48]), it was seen that for these power series s_k could be chosen so that $|r_0^{(k)}v_0^{(k)}| > 10^{-2}$.

7. Bounds on the relative error in the computed Padé approximants and inverses of Hankel matrices.

We now consider the problem of obtaining bounds on $(\delta U^{(k)}, \delta V^{(k)})$ and $(\delta P^{(k)}, \delta Q^{(k)})$. For the moment, we dispense with the superscripts on the quantities being discussed in order to simplify the notation. The superscripts will reappear when they are needed again for clarity. Thus, from (10), assume that Algorithm NPADE (after k iterations) gives $\mathcal{P}$ such that

$$\mathcal{P}\mathcal{A} = \mathcal{D} + O(z^{2n}). \tag{67}$$

Then from (4) and (67), we have

$$\begin{pmatrix} a_0 & \cdots & a_{n-1} \\ \vdots & & \vdots \\ a_{n-1} & \cdots & a_{2n-2} \end{pmatrix}\begin{pmatrix} v_n \\ \vdots \\ v_1 \end{pmatrix} = -v_0\begin{pmatrix} a_n \\ \vdots \\ a_{2n-1} \end{pmatrix} + \begin{pmatrix} \delta w_n \\ \vdots \\ \delta w_{2n-1} \end{pmatrix}.$$

The matrix on the left-hand side is the Hankel matrix $H_{n-1,n}$. Since $r_0v_0 \neq 0$ at this step, the Hankel matrix is invertible (Labahn and Cabay [38]) and we can easily obtain an expression for δV in terms of δW. Then (3), which defines $U(z)$ in terms of $V(z)$, can be used to obtain δU. In a like fashion, we can obtain expressions for $(\delta P, \delta Q)$ in terms of δR. Thus, we can express the error in the computed solutions if we have an expression for $H_{n-1,n}^{-1}$.

Now, dropping the superscripts, (47) can be rewritten as

$$QU - VP = \vartheta(z) + v_0r_0z^{2n}, \tag{68}$$

where

$$\vartheta(z) = (V\delta R - Q\delta W) \pmod{z^{2n+1}}. \tag{69}$$

With these relations, we can prove the following.

THEOREM 7.1. *Let $\mathcal{P}$ be given by* (67). *Then*

$$H_{n-1,n}\left(\frac{1}{r_0v_0}\tilde{H}_{n-1,n}\right) = I + \frac{1}{r_0v_0}E, \tag{70}$$

where

$$\tilde{H}_{n-1,n} = \begin{pmatrix} v_{n-1} & \cdots & v_0 \\ \vdots & \cdot^{\cdot^{\cdot}} & \\ v_0 & & 0 \end{pmatrix} \begin{pmatrix} q_{n+1} & \cdots & q_2 \\ & \ddots & \vdots \\ 0 & & q_{n+1} \end{pmatrix}$$

$$(71) \qquad - \begin{pmatrix} q_n & \cdots & q_2 & 0 \\ \vdots & \cdot^{\cdot^{\cdot}} & \cdot^{\cdot^{\cdot}} & \vdots \\ q_2 & \cdot^{\cdot^{\cdot}} & & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix} \begin{pmatrix} v_n & \cdots & v_1 \\ & \ddots & \vdots \\ 0 & & v_n \end{pmatrix},$$

and

$$(72) \qquad E = [\vartheta_{2n-2+i-j}] + [\delta w_{n-1+i-j}][q_{n+1+i-j}] - [\delta r_{n+i-j}][v_{n+i-j}].$$

In the above theorem and the proof that now follows, we have used $[f(i,j)]$ to denote the matrix whose (i,j) entry is $f(i,j)$. Moreover, an entry is taken to be zero if the resulting subscript is negative[14] or else exceeds the degree of the corresponding polynomial.

Proof. We provide an outline of the proof of this theorem. The details for the case where $\delta R^{(k)} = \delta W^{(k)} = 0$ can be found in Labahn, Choi, and Cabay [39] and the general case in Meleshko [47].

From (47), letting $1 \leq i,j \leq n$, we have

$$[u_{n-1+i-j}][q_{n+1+i-j}] - [p_{n+i-j}][v_{n+i-j}] = r_0 v_0 I + [\vartheta_{2n+i-j}].$$

Similarly,

$$[v_{2n+1-i-j}][q_{n+1+i-j}] - [q_{2n+2-i-j}][v_{n+i-j}] = 0.$$

Then, by equating the coefficients of $z^{n-1+i-j}$ in (45), it can be shown that

$$H_{n-1,n}[v_{n+1-i-j}] = [u_{n+i-j-1}] - H_{-1,n}[v_{2n+1-i-j}] + [\delta w_{n+i-j-1}].$$

Likewise, equating the coefficients of z^{n+i-j} in (44), yields

$$H_{n-1,n}[q_{n+2-i-j}] = [p_{n+i-j}] - H_{-1,n}[q_{2n+2-i-j}] + [\delta r_{n+i-j}].$$

The result then follows by applying the above relationships to the product $H_{n-1,n}\tilde{H}_{n-1,n}$. □

Equation (71) provides a formula for the inverse of $(r_0 v_0)^{-1} H_{n-1,n}$. Moreover, by reordering the unknowns (see (2)), a similar formula for the inverse of a Toeplitz matrix can be obtained. Since $\tilde{H}_{n-1,n} b$, for any vector b, can be obtained by a sequence of polynomial multiplications, the ability to compute $\tilde{H}_{n-1,n}$ in time $O(n^2)$ using Algorithm NPADE also provides a fast method for solving Hankel and Toeplitz systems. For the exact case, the formula for expressing the inverses of Toeplitz and Hankel matrices in terms of Padé approximants can be found in Labahn, Choi, and Cabay [39]. Also, for the exact case formula (71) can be derived from formulae given

[14] Also recall that $q_0 = q_1 = p_0 = p_1 = 0$.

by Gohberg and Semencul [29] and Iohvidov [35] which express the inverse of $H_{n-1,n}$ entirely in terms of the first and last columns of the inverse.

Intuitively, if $\| (1/r_0v_0)E \|$ is small, then $(1/r_0v_0)\tilde{H}_{n-1,n}$ is close to the inverse of $H_{n-1,n}$ in a certain sense. We make this idea more precise now.

LEMMA 7.2. *If* $\| (1/r_0v_0)E \| < 1$, *then*

$$H_{n-1,n}^{-1} = \left(\frac{1}{r_0v_0}\tilde{H}_{n-1,n}\right)\left(I + \frac{1}{r_0v_0}E\right)^{-1} \tag{73}$$

and

$$\left\|\left(I + \frac{1}{r_0v_0}E\right)^{-1}\right\| \le \frac{1}{1 - \| \frac{1}{r_0v_0}E \|}. \tag{74}$$

Proof. Since

$$\left\|\frac{1}{r_0v_0}E\right\| < 1, \quad \text{then } I + \frac{1}{r_0v_0}E$$

is invertible (Stewart [52, p. 187]), so that (73) follows from (70). A proof of (74) can be found in Stewart [52, p. 187]. □

Letting

$$J_n = \begin{pmatrix} 0 & & a_0 \\ & \cdot^{\cdot^{\cdot}} & \vdots \\ a_0 & \cdots & a_n \end{pmatrix},$$

the order conditions (44) and (45) give us

$$H_{n-1,n}\begin{pmatrix} v_n \\ \vdots \\ v_1 \end{pmatrix} = -v_0\begin{pmatrix} a_n \\ \vdots \\ a_{2n-1} \end{pmatrix} + \begin{pmatrix} \delta w_n \\ \vdots \\ \delta w_{2n-1} \end{pmatrix}, \tag{75}$$

$$J_{n-1}\begin{pmatrix} v_{n-1} \\ \vdots \\ v_0 \end{pmatrix} = \begin{pmatrix} u_0 \\ \vdots \\ u_{n-1} \end{pmatrix} + \begin{pmatrix} \delta w_0 \\ \vdots \\ \delta w_{n-1} \end{pmatrix}, \tag{76}$$

$$H_{n-1,n}\begin{pmatrix} q_{n+1} \\ \vdots \\ q_2 \end{pmatrix} = r_0e_n + \begin{pmatrix} \delta r_{n+1} \\ \vdots \\ \delta r_{2n-1} \\ 0 \end{pmatrix}, \tag{77}$$

$$J_{n-2}\begin{pmatrix} q_n \\ \vdots \\ q_2 \end{pmatrix} = \begin{pmatrix} p_2 \\ \vdots \\ p_n \end{pmatrix} + \begin{pmatrix} \delta r_2 \\ \vdots \\ \delta r_n \end{pmatrix}, \tag{78}$$

keeping in mind that the constant and linear terms in P, Q, and δR are zero. From (75), (76), (77), and (78), and the algebraic formulation for the exact quantities, it is easy to see that the following is true.

LEMMA 7.3.

$$\begin{pmatrix} \delta v_n \\ \vdots \\ \delta v_0 \end{pmatrix} = \left(H_{n-1,n}^{-1} \begin{pmatrix} \delta w_n \\ \vdots \\ \delta w_{2n-1} \\ 0 \end{pmatrix} \right),$$

$$\begin{pmatrix} \delta u_0 \\ \vdots \\ \delta u_{n-1} \end{pmatrix} = J_{n-1} \begin{pmatrix} \delta v_{n-1} \\ \vdots \\ \delta v_0 \end{pmatrix} - \begin{pmatrix} \delta w_0 \\ \vdots \\ \delta w_{n-1} \end{pmatrix},$$

$$\begin{pmatrix} \delta q_{n+1} \\ \vdots \\ \delta q_2 \end{pmatrix} = H_{n-1,n}^{-1} \begin{pmatrix} \delta r_{n+1} \\ \vdots \\ \delta r_{2n-1} \\ 0 \end{pmatrix}, \text{ and}$$

$$\begin{pmatrix} \delta p_2 \\ \vdots \\ \delta p_n \end{pmatrix} = J_{n-2} \begin{pmatrix} \delta q_n \\ \vdots \\ \delta q_2 \end{pmatrix} - \begin{pmatrix} \delta r_2 \\ \vdots \\ \delta r_n \end{pmatrix}.$$

Immediately, from the preceding lemma, we have the following corollary.

COROLLARY 7.4.

$$\| \delta P \| + \| \delta Q \| \leq (n \| H_{n-1,n}^{-1} \| + 1) \| \delta R \| \tag{79}$$

and

$$\| \delta U \| + \| \delta V \| \leq ((n+1) \| H_{n-1,n}^{-1} \| + 1) \| \delta W \|. \tag{80}$$

We have bounds for $\| \delta R \|$ and $\| \delta W \|$; the following gives us a bound for $\| H_{n-1,n}^{-1} \|$.

LEMMA 7.5. *If*

$$22(n+1)^2(\| \delta R \| + \| \delta W \| + 5(n+1)\mu) \leq |r_0 v_0|, \tag{81}$$

then

$$\| H_{n-1,n}^{-1} \| \leq \frac{2.2n^2}{|r_0 v_0|}. \tag{82}$$

Proof. From (70), we have that

$$H_{n-1,n} \left(\frac{1}{r_0 v_0} \tilde{H}_n \right) = I + \frac{1}{r_0 v_0} E,$$

where, using (71), (21), and (22),

$$\| \tilde{H}_n \| \leq 2n^2, \tag{83}$$

and, from (72),

$$\left\| \frac{1}{r_0 v_0} E \right\| \leq \frac{2n(n+1)(\| \delta R \| + \| \delta W \|)}{|r_0 v_0|}. \tag{84}$$

Applying (81), which is the same assumption as used in Lemma 6.7, we have that

$$\left\| \frac{1}{r_0 v_0} E \right\| \le \frac{1}{11} < 1. \tag{85}$$

Since (85) holds, Lemma 7.2 and (83) give

$$H_{n-1,n}^{-1} = \left(\frac{\tilde{H}_n}{r_0 v_0} \right) \left(I + \frac{1}{r_0 v_0} E \right)^{-1}$$

and

$$\begin{aligned} \| H_{n-1,n}^{-1} \| &\le \left(\frac{\| \tilde{H}_n \|}{|r_0 v_0|} \right) \left\| \left(I + \frac{1}{r_0 v_0} E \right)^{-1} \right\| \\ &\le \left(\frac{2n^2}{|r_0 v_0|} \right) (1.1) \end{aligned}$$

and the result follows. □

With this last lemma, we can now state our principal theorem.

THEOREM 7.6. *If*

$$22(n_l+1)^2(\| \delta R^{(l)} \| + \| \delta W^{(l)} \| + 5(n_l+1)\mu) \le |r_0^{(l)} v_0^{(l)}| \tag{86}$$

holds for $0 \le l \le k$, *then*

$$\begin{aligned} &\begin{pmatrix} \| \delta P^{(k)} \| + \| \delta Q^{(k)} \| \\ \| \delta U^{(k)} \| + \| \delta V^{(k)} \| \end{pmatrix} \\ &\quad \le \frac{2.2(n_k+1)^3}{|\sigma^{(k)}|} \left(\frac{\Gamma_0^{(k-1)}}{|\sigma^{(k-1)}|} + \sum_{l=0}^{k-2} \frac{\Gamma_1^{(l)}}{|\sigma^{(l)} \sigma^{(l+1)}|} \right) \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix} + O(\mu^2) \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \end{aligned} \tag{87}$$

where Γ_0 *and* Γ_1 *are as defined in* (13) *and* (14).

Proof. This is immediate from Theorem 6.11, Corollary 7.4, and Lemma 7.5, with the relaxation of some of the inequalities even further. □

Using the fact that $1/|\sigma^{(i)}| < 1/\varepsilon$, we obtain the stability Theorem 5.1 of §5.

Since $\| P^{(k)} \| + \| Q^{(k)} \| = \| U^{(k)} \| + \| V^{(k)} \| = 1$, Theorem 7.6 gives a bound on the relative error in the computed Padé approximants. In Theorem 7.6, $\sigma^{(k)}$ appears in just the first factor of the term linear in μ in (87); this factor comes from the condition number of the Hankel matrix that corresponds to the Padé approximant that is finally output by Algorithm NPADE. The second factor in the linear term in (87) comes from the bound on $\mathcal{D}^{(k)}$ and does not involve $\sigma^{(k)}$. Thus, if the algorithm computes its last Padé approximant using an unstable block (i.e., for $|\sigma^{(k)}|$ small), the residual of the final Padé approximant will still be small as $|\sigma^{(0)}|, \ldots, |\sigma^{(k-1)}|$ will all exceed ε. The relative error in $(U^{(k)}, V^{(k)})$ may be large, though, as this quantity does involve $|\sigma^{(k)}|$.

8. Experimental results. Numerical experiments have been performed to compare the analysis of the algorithm with its practice. We present a summary of the conclusions here; details appear in Meleshko and Cabay [48]. Table 1 gives the results of one typical experiment for a power series $A(z)$ whose coefficients are random and uniformly distributed between -1 and 1. In this table, the error in the order conditions

TABLE 1
Experimental results for the error in the computed Padé approximants.

		$\Vert \delta R^{(k)} \Vert + \Vert \delta W^{(k)} \Vert$		$\Vert \delta U^{(k)} \Vert + \Vert \delta V^{(k)} \Vert$	
n_k	$\sigma^{(k)}$	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-12}$	$\varepsilon = 10^{-2}$	$\varepsilon = 10^{-12}$
2	4.223e−01	5.898e−17	5.898e−17	5.5511e−17	5.5511e−17
3	1.440e−01	1.110e−16	1.110e−16	8.3267e−17	8.3267e−17
4	1.240e−01	1.665e−16	1.665e−16	1.9082e−16	1.9082e−16
5	3.099e−02	1.943e−16	1.943e−16	2.2204e−16	2.2204e−16
6	8.118e−02	6.939e−16	6.939e−16	4.7184e−16	4.7184e−16
7	1.894e−01	1.776e−15	1.776e−15	1.1796e−15	1.1796e−15
8	2.628e−01	2.998e−15	2.998e−15	2.3176e−15	2.3176e−15
9	5.165e−08	Unstable	3.646e−15	Unstable	4.2466e−15
10	3.172e−01	4.111e−15	7.231e−10	3.8580e−15	1.1035e−09
11	1.951e−09	Unstable	1.446e−09	Unstable	9.1274e−10
12	2.927e−01	4.954e−15	2.628e−08	8.0908e−15	3.1795e−08
13	3.095e−06	Unstable	5.114e−08	Unstable	4.1980e−08
14	2.447e−01	5.931e−15	6.7887e−08	6.8556e−15	6.0450e−08
15	6.627e−04	Unstable	8.571e−08	Unstable	4.2025e−07
16	2.396e−07	Unstable	8.682e−08	Unstable	1.0281e−07
17	2.975e−02	4.732e−15	9.160e−08	4.1078e−15	6.1857e−08
18	2.145e−01	5.039e−15	8.370e−08	5.8703e−15	5.2384e−08
19	1.293e−01	6.773e−15	6.949e−08	1.1754e−14	4.5629e−08
20	6.928e−03	Unstable	7.003e−08	Unstable	1.9954e−07
21	6.825e−06	Unstable	7.200e−08	Unstable	2.4984e−07
22	4.860e−03	Unstable	7.274e−08	Unstable	1.4915e−07
23	1.030e−01	7.979e−15	8.379e−08	6.8279e−15	7.0215e−08
24	1.492e−01	7.310e−15	1.291e−07	5.3291e−15	3.1958e−07
25	7.601e−02	7.690e−15	1.656e−07	7.7716e−15	1.5034e−07
26	2.801e−02	9.094e−15	1.493e−07	6.0785e−15	1.3825e−07
27	1.222e−01	8.558e−15	1.305e−07	4.3923e−15	7.5392e−08
28	1.724e−01	7.265e−15	1.241e−07	5.5789e−15	6.5606e−08
29	4.828e−02	7.105e−15	1.166e−07	2.9657e−14	1.1872e−07
30	5.095e−02	7.272e−15	1.321e−07	6.4115e−15	1.8561e−07

and the computed Padé approximants are given for two values of the stability parameter ε. The value $\varepsilon = 10^{-2}$ indicates a willingness to accept only Sylvester matrices S_{n_l} with condition numbers less than $1/\varepsilon = 10^2$, approximately. Sylvester matrices not satisfying this criterion are assumed to lie in an unstable block and are skipped over. The value $\varepsilon = 10^{-12}$ permits a much greater tolerance for ill-conditioning and results in an expected deterioration in the accuracy.

It was observed that the large constants and powers of n_l and s_l that occur in the bounds derived above were not manifested in the experiments. Also $\mathcal{D}$ depends on $|\sigma^{(l)}|$ and not $|\sigma^{(l)}\sigma^{(l+1)}|$ and the overall error was inversely proportional to the smallest $|\sigma^{(l)}|$ encountered. An operational bound on the error in the order condition would be

$$(88) \qquad |\mathcal{D}| \leq \left(\sum_{l=0}^{k-1} \frac{n_l M \rho_l}{|\sigma^{(l)}|} \right) \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix} + O(\mu^2) \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where M is a moderate (approximately $O(10)$) constant.

Then, in terms of the Padé approximant, (88) yields

$$(89) \qquad \Vert \delta U \Vert + \Vert \delta V \Vert \leq \frac{1}{|\sigma^{(k)}|} \left(\sum_{l=0}^{k-1} \frac{n_l M \rho_l}{|\sigma^{(l)}|} \right) \mu \begin{pmatrix} 1 \\ 1 \end{pmatrix} + O(\mu^2) \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

As an a posteriori estimate, we have

$$\| \delta U \| + \| \delta V \| \le \frac{K}{\varepsilon}(\| \delta R \| + \| \delta W \|), \tag{90}$$

where K is a modest (approximately $O(10)$) constant.

9. Conclusion. Algorithm NPADE is a weakly stable algorithm for computing Padé approximants for the general class of power series. Except in exceptionally degenerate cases, Algorithm NPADE is an $O(n^2)$ fast algorithm. (By this we mean that s_i is bounded by some modest constant; in a series of experiments with random power series and $\varepsilon = 10^{-2}$ reported in Meleshko and Cabay [48], the stepsize never exceeded eight and was usually less than four.) Also, once the Padé approximants are obtained, one can use Q and V to obtain the inverse of the Hankel matrix $H_{n-1,n}$.

The relation that enabled so strong a conclusion to be drawn was (47), namely,

$$QU - VP = \vartheta(z) + r_0 v_0 z^{m+n-1},$$

where $\vartheta(z)$ depends on δR and δW, the errors in the order conditions. This relation also permitted the derivation of the inverse of a Hankel matrix in terms of a Padé approximant.

An important question is how ε should be chosen. By replacing the σ by the stability parameter ε, equation (89) suggests that the loss of accuracy in the computed $(n-1, n)$ Padé approximant is approximately $\log_{10}(n\varepsilon^{-2})$ digits. This provides a rule of thumb for choosing ε, but it is based on an a priori estimate of the error and so may be somewhat pessimistic in the ε it suggests.

As remarked at the end of §4, Algorithm NPADE is equivalent to the algorithm of Cabay and Choi [16] if exact arithmetic is used and ε is 0. In [16], Cabay and Choi give a superfast $O(n \log^2 n)$ version of their algorithm. It is an open question as to whether or not a larger stepsize and fast polynomial multiplication can be exploited by Algorithm NPADE to produce a stable superfast variant. Also in [16], Cabay and Choi exhibit a duality between their algorithm and the Euclidean algorithm. An interesting question is what this duality may mean in a numerical setting. (See, for example, the discussion in [51].)

One last question is whether or not the stability result can be strengthened. In the worst case, Algorithm NPADE is numerically equivalent to Gaussian elimination with partial pivoting, which is a stable algorithm.

REFERENCES

[1] G. S. Ammar and W. B. Gragg, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[2] ———, *Numerical experience with a superfast real Toeplitz solver*, Linear Algebra Appl., 121 (1989), pp. 185–206.

[3] A. Antoulas, J. Ball, J. Kang, and J. Willems, *On the solution of the minimal rational interpolation problem*, Linear Algebra Appl., 137/138 (1990), pp. 511–573.

[4] G. Baker and P. Graves-Morris, *Padé Approximants*, Encyclopedia of Mathematics, Vol. 13, Addison-Wesley, Reading, MA, 1981.

[5] ———, *Padé Approximants*, Encyclopedia of Mathematics, Vol. 14, Addison-Wesley, Reading, MA, 1981.

[6] E. H. Bareiss, *Numerical solution of linear equations with Toeplitz and vector matrices*, Numer. Math., 13 (1969), pp. 404–424.

[7] B. Beckermann, *A reliable method for computing M-Padé approximants on arbitrary staircases*, J. Comput. Appl. Math., 39 (1992), pp. 295–313.

[8] E. R. BERLEKAMP, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

[9] R. R. BITMEAD AND B. D. O. ANDERSON, *Asymptotically fast solutions of Toeplitz and related systems of linear equations*, Linear Algebra Appl., 34 (1980), pp. 103–116.

[10] R. BRENT, F. GUSTAVSON, AND D. YUN, *Fast solution of Toeplitz systems of equations and computation of Padé approximants*, J. Algorithms, 1 (1980), pp. 259–295.

[11] A. BULTHEEL, *Recursive algorithms for nonnormal Padé tables*, SIAM J. Appl. Math., 39 (1980), pp. 106–118.

[12] A. BULTHEEL AND M. VAN BAREL, *Padé techniques for model reduction in linear system theory: A survey*, J. Comput. Appl. Math., 14 (1986), pp. 401–438.

[13] A. BULTHEEL AND L. WUYTACK, *Stability of numerical methods for computing Padé approximants*, in Approximation Theory III, 1980, Proc. Conf. Univ. of Texas, Austin, A. Magnus, ed., Academic Press, New York, pp. 267–272.

[14] J. R. BUNCH, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.

[15] ———, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl., 88/89 (1987), pp. 49–66.

[16] S. CABAY AND D. CHOI, *Algebraic computations of scaled Padé fractions*, SIAM J. Comput., 15 (1986), pp. 243–270.

[17] S. CABAY AND A. JONES, *Experiments with a stable algorithm for computing Padé–Hermite approximants*, Tech. Rep., Dept. of Computing Science, University of Alberta, Edmonton, Canada, to appear.

[18] S. CABAY AND P. KOSSOWSKI, *Power series remainder sequences and Padé fractions over an integral domain*, J. Symbolic Comput., 10 (1990), pp. 139–163.

[19] S. CABAY AND G. LABAHN, *A superfast algorithm for multi-dimensional Padé approximants*, J. Numer. Algorithms, 2 (1992), pp. 201–224.

[20] S. CABAY, G. LABAHN, AND B. BECKERMANN, *On the theory and computation of non-perfect Padé-Hermite approximants*, J. Comput. Appl. Math., 39 (1992), pp. 295–313.

[21] G. CLAESSENS, *A new look at the Padé table and the different methods for computing its elements*, J. Comput. Appl. Math., 1 (1975), pp. 141–153.

[22] G. CLAESSENS AND L. WUYTACK, *On the computation of non-normal Padé approximants*, J. Comput. Appl. Math., 5 (1979), pp. 282–289.

[23] G. CYBENKO, *The numerical stability of the Levinson-Durbin algorithm for Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 303–319.

[24] F. DE HOOG, *A new algorithm for solving Toeplitz systems of equations*, Linear Algebra Appl., 88/89 (1987), pp. 123–138.

[25] P. DELSARTE AND Y. V. GENIN, *The split Levinson algorithm*, IEEE Trans. ASSP, 34 (1986), pp. 470–478.

[26] G. E. FORSYTHE AND C. B. MOLER, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[27] G. FROBENIUS, *Relationen zwischen den näherungsbrüchen von potenzreihen*, J. für Math., 30 (1881), pp. 1–17.

[28] K. O. GEDDES, *Symbolic computation of Padé approximants*, ACM Trans. Math. Software, 5 (1979), pp. 218–233.

[29] I. C. GOHBERG AND A. A. SEMENCUL, *On the inversion of finite Toeplitz matrices and their continuous analogs*, Mat. Issled., 2 (1972), pp. 201–233.

[30] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[31] W. B. GRAGG, *The Padé table and its relation to certain algorithms of numerical analysis*, SIAM Rev., 14 (1972), pp. 1–62.

[32] W. B. GRAGG, F. G. GUSTAVSON, D. D. WARNER, AND D. Y. Y. YUN, *On fast computation of superdiagonal Padé fractions*, Math. Programming Stud., 18 (1982), pp. 39–42.

[33] M. H. GUTKNECHT, personal communication.

[34] N. J. HIGHAM AND D. J. HIGHAM, *Large growth factors in Gaussian elimination with pivoting*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 155–164.

[35] I. IOHVIDOV, *Hankel and Toeplitz Matrices and Forms—Algebraic Theory*, Birkhauser, Boston, 1982.

[36] J. R. JAIN, *An efficient algorithm for a large Toeplitz set of linear equations*, IEEE Trans. Acoust. Speech Signal Proces., AASP-27 (1979), pp. 612–615.

[37] T. KAILATH, *A view of three decades of linear filtering theory*, IEEE Trans. Inform. Theory, 20 (1974), pp. 146–181.

[38] G. LABAHN AND S. CABAY, *Matrix Padé fractions and their computation*, SIAM J. Comput., 18 (1989), pp. 639–657.

[39] G. Labahn, D. K. Choi, and S. Cabay, *The inverses of block Hankel and block Toeplitz matrices*, SIAM J. Comput., 19 (1990), pp. 98–123.

[40] N. Levinson, *The Weiner RMS error criterion in filter design and prediction*, Stud. Appl. Math., 25 (1947), pp. 261–278.

[41] F. T. Luk and S. Qiao, *A fast but unstable orthogonal triangularization technique for Toeplitz matrices*, Linear Algebra Appl., 88/89 (1987), pp. 495–506.

[42] Y. L. Luke, *Computations of coefficients in the polynomials of Padé approximations by solving systems of linear equations*, J. Comput. Appl. Math., 6 (1980), pp. 213–218.

[43] ———, *A note on evaluation of coefficients in the polynomials of Padé approximants by solving systems of linear equations*, J. Comput. Appl. Math., 8 (1982), pp. 93–99.

[44] J. Makhoul, *Linear prediction: A tutorial review*, Proc. IEEE, 63 (1975), pp. 561–580.

[45] J. L. Massey, *Shift register synthesis and bch decoding*, IEEE Trans. Inform. Theory, IT-15 (1969), pp. 122–177.

[46] R. Meleshko, *A stable algorithm for the computation of Padé approximants*, Tech. Report TR 90-19, Dept. of Computing Science, University of Alberta, Edmonton, Canada, 1990.

[47] ———, *A stable algorithm for the computation of Padé approximants*, Ph.D. thesis, Dept. of Computing Science, University of Alberta, Edmonton, Canada, 1990.

[48] R. Meleshko and S. Cabay, *On computing Padé approximants quickly and accurately*, Congr. Numer., 80 (1991), pp. 245–255.

[49] H. Padé, *Sur la représentation approchés d'une fonction par des fractions rationelles*, Ann. Ec. Norm., Sup., 3 (1892), pp. 1–93.

[50] J. Rissanen, *Solution of linear equations with Hankel and Toeplitz matrices*, Numer. Math., 22 (1974), pp. 361–366.

[51] T. W. Sederberg, *Improperly parametrized rational curves*, Computer Aided Geom. Des., 3 (1986), pp. 67–75.

[52] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[53] D. R. Sweet, *Fast Toeplitz orthogonalization*, Numer. Math., 43 (1984), pp. 1–21.

[54] L. N. Trefethen and R. S. Schreiber, *Average-case stability of Gaussian elimination*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 335–360.

[55] W. Trench, *An algorithm for the inversion of finite Hankel matrices*, J. Soc. Indust. Appl. Math., 13 (1965), pp. 1102–1107.

[56] M. Van Barel and A. Bultheel, *A new approach to the rational interpolation problem*, J. Comput. Appl. Math., 32 (1990), pp. 281–289.

[57] ———, *A new formal approach to the rational interpolation problem*, Numer. Math., 62 (1992), pp. 87–122.

[58] G. Walker, *On periodicity in series of related terms*, Proc. Roy. Soc. London, Ser. A, 131A (1931), p. 518.

[59] J. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1963.

[60] ———, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

[61] J. H. Wilkinson, *Modern error analysis*, SIAM Rev., 13 (1971), pp. 548–568.

[62] P. Wynn, *On the convergence and stability of the epsilon algorithm*, SIAM J. Numer. Anal., 3 (1966), pp. 91–122.

[63] G. U. Yule, *On a method of investigating periodicities in disturbed series, with special reference to Wolfe's sunspot numbers*, Philos. Trans. Roy. Soc. London, Ser. A., 226A (1927), pp. 267–298.

[64] S. Zohar, *Toeplitz matrix inversion: The algorithm of W.F. Trench*, J. Assoc. Comput. Mach., 16 (1969), pp. 592–601.

[65] ———, *The solution of a Toeplitz set of linear equations*, J. Assoc. Comput. Mach., 21 (1974), pp. 272–276.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 766–775, July 1993

NESTED DISSECTION FOR SPARSE NULLSPACE BASES*

JULIO M. STERN[†] AND STEPHEN A. VAVASIS[‡]

Abstract. The authors propose a nested dissection approach to finding a fundamental cycle basis in a planar graph. The cycle basis corresponds to a fundamental nullspace basis of the adjacency matrix. This problem is meant to model sparse nullspace basis computations occurring in a variety of settings. An $O(n^{3/2})$ bound is achieved on the nullspace basis size (i.e., the number of nonzero entries in the basis), and an $O(n \log n)$ bound on the size in the special case of grid graphs.

Key words. sparse, nullspace, basis, force method, nested dissection

AMS subject classifications. 65F05, 65F50

1. Fundamental nullspace bases and graphs.

Let A be an $n \times m$ matrix with $m \geq n$, and let its rank be r. An important problem in scientific computation is to produce a basis for the nullspace of A. In other words, we want to compute an $m \times (m-r)$ matrix V of rank $m-r$ such that $AV = 0$. The columns of V satisfy the equation $Av = 0$; such a vector v is usually called a *null vector*.

The nullspace basis problem arises in the following context, as explained further by Strang [15]. Solving the square nonsingular linear system

$$\begin{pmatrix} H & -A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{1}$$

is the key step for structural analysis, the finite element method, optimization problems, and electrical network analysis. In these applications, H is diagonal or block diagonal and symmetric positive definite, and A has full row rank. Usually, H encodes element stiffnesses or resistances, and A encodes geometry and connectivity among elements. Two common methods to solve this include the "displacement" method and the "force" method (see Kaneko, Lawo, and Thierauf [8]), terminology arising in the structural analysis application. The displacement method involves substituting the first block of equations $x = H^{-1}(A^Ty + f)$ into the second formula $Ax = g$, arriving at a symmetric positive definite linear system with y as the only unknown. The resulting coefficient matrix $AH^{-1}A^T$ is called the "assembled stiffness matrix" in the context of finite elements.

The force method instead attempts to eliminate y and to solve for x. Let V be a nullspace basis for A. Let x_0 be any solution for $Ax_0 = g$ (typically there is a simple application-dependent method for finding some solution to the underdetermined system $Ax = g$). Then we know that the vector x, which is the first solution component to (1), satisfies $A(x - x_0) = 0$, i.e., $x - x_0 = Vq$ for some vector

* Received by the editors May 26, 1991; accepted for publication (in revised form) December 27, 1991. This work was supported in part by the Applied Mathematical Sciences Program (KC-04-02) of the Office of Energy Research of the U.S. Department of Energy under grant DE-FG02-86ER25013.A000, the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell, and by a National Science Foundation Presidential Young Investigator award. The second author's work was supported in part by grant 87/0540-9 from FAPESP, Fundação de Amparo à Pesquisa do Estado de São Paulo, Brasil.

[†] School of Operations Research, Cornell University, Ithaca, New York, 14853. Present address, Departamento de Ciência da Computação do Instituto de Matemática e Estatística da Universidade de São Paulo, 05508 São Paulo, Brasil (jstern@ime.usp.br).

[‡] Department of Computer Science, Cornell University, Ithaca, New York 14853 (vavasis@cs.cornell.edu).

q. Then we can substitute $x = Vq + x_0$ in the equation $Hx - A^Ty = f$ to obtain $H(Vq+x_0)-A^Ty = f$. Now multiply on the left by V^T, using the fact that $V^TA^T = 0$, obtaining $V^TH(Vq - x_0) = V^Tf$. This is a linear system that may be solved for q, hence obtaining x. The force method can be advantageous in circumstances involving fixed geometry but changing structural properties.

If A is sparse (as usually happens in applications), then we might hope to compute a nullspace basis V that is itself sparse. This has been a topic of recent interest in the literature; it has been addressed by Coleman and Pothen [3], Gilbert and Heath [6], and Plemmons and White [11]. In general there is no reason to believe that sparsity in A implies sparsity in V: it is necessary that A have some additional structure.

None of these earlier works are able to establish bounds on the number of nonzero entries in V. In this report we propose an algorithm that computes a nullspace basis with an asymptotic bound on the number of nonzero entries in V. In order to establish bounds, we focus attention on a simple model problem that captures the features of many real problems.

If V contains an embedded $(m-r) \times (m-r)$ identity matrix, the basis is said to be *fundamental.* In other words, there exist permutation matrices P and Q such that

$$PVQ = \begin{pmatrix} I \\ C \end{pmatrix}.$$

Fundamental nullspace bases are especially useful in certain applications because they lead to simplified algorithms. In particular, the rows of V not in the identity matrix correspond to columns of A forming a basis of the span of A, and the null vectors give equations that express each nonbasic column of A in terms of the basis columns. In this report we focus on the fundamental nullspace basis problem.

D. Rose, in a private communication, has observed that the problem of finding a fundamental nullspace basis of A corresponds to a certain symmetric elimination order applied to (1). This implies that both the force method (in the context of a fundamental basis) and displacement method can be put into a more general framework of elimination orderings. We do not pursue this generalization here.

The particular model problem we have selected is the case that A is the node-edge incidence matrix of an undirected graph. This is a matrix with n rows, one for each vertex of the graph, and m columns, one for each edge. Each column has two nonzero entries in the positions corresponding to the endpoints of the edge. These two entries are one $+1$ and one -1 in each column in arbitrary order. (In some applications, the graph is actually directed, in which case the signs of the entries indicate edge orientation. However, the choice of orientation has no effect on the construction of the nullspace basis since multiplying a column by -1 does not change its span.)

A nullspace basis for such a matrix has a natural combinatorial interpretation. In particular, the nonzero entries of a particular column of V corresponds to a closed walk in the graph. A fundamental nullspace basis for a connected graph corresponds to the identification of a spanning tree T of the graph. A null vector corresponds to the unique cycle formed by edges of T in conjunction with a single edge of $G-T$. See Welsh [17] for more information.

Thus, the problem of finding a fundamental nullspace basis with suitable properties is reduced to the problem of finding the right spanning tree T of G. We propose an algorithm reminiscent of nested dissection for finding this tree T. Nested dissection, a technique due to George [5], finds a *separator* of the graph and then recursively works with subgraphs. Here, a *separator* refers to a small set of nodes whose removal disconnects the graph into pieces of size at most $2n/3$ nodes. The exact definition is contained in the next section. Unlike traditional nested dissection, which requires no further properties of the separator, we will also need the separator nodes to form a cycle.

Such separators are found in planar graphs that have a bound on the maximum face size, a result due to Miller [9]. See the next section for the theorem.

The existence of any kind of separator automatically means that A can be partitioned in *row block angular form* (RBAF) if the rows corresponding to separator vertices are numbered last. An example of block angular form is

$$\begin{pmatrix} B_1 & & & & & \\ & B_2 & & & & \\ & & B_3 & & & \\ & & & \ddots & & \\ & & & & B_k & \\ S_1 & S_2 & S_3 & \cdots & S_k & S_{k+1} \end{pmatrix}. \tag{2}$$

In this regard, our approach is most reminiscent of Plemmons and White. They work with *column block angular form* (CBAF) and do not attempt to derive bounds on the number of nonzero entries in the basis.

It may seem that we are speaking of a very restricted class of matrices by limiting attention to node-edge adjacency matrices of planar graphs, but this is a model for the kind of matrices that occur in practice. For example, optimization problems with flow constraints have constraint equations in the form of a graph, and the graph is typically planar or nearly planar. As another example, Pothen [12] shows that for certain kinds of structural problems, the nullspace basis of the structural equilibrium matrix A can be entirely deduced from a cycle basis for the underlying graph.

The remainder of this paper is organized as follows. In §2 we give an algorithm that achieves an $O(n^{3/2})$ bound on the size of the nullspace. An added bonus of our algorithm is that the resulting basis has structure useful for parallelism. In §3 we specialize our algorithm to the case of a grid graph, achieving an $O(n \log n)$ algorithm. Finally, in §4 we discuss how to generalize our ideas to other kinds of matrices.

In many applications where A is not explicitly given in RBAF, we can permute rows and columns to put PAQ in RBAF. Finding the permutation matrices P and Q that minimize the number of residual rows or columns, while producing diagonal blocks of approximately the same size, is in general a very difficult combinatorial problem to be solved exactly. See work by Stern [14].

2. A tree from nested simple cycle separators. In a graph $G = (V, E)$ with $|V| = n$, a set $S \subset V$ is called an (α, β)*-separator* if and only if

1. Set S contains at most $\beta\sqrt{n}$ vertices, and
2. Graph $G - S$ has no connected component with more than αn vertices.

For this section we need the following theorem from Miller [9].

THEOREM 2.1. *Let G be a biconnected planar graph with n vertices, m edges, and maximal face size φ. Then there is an (α, β)-SCS (simple cycle separator) S, with $\alpha = 2/3$ and $\beta = 2\sqrt{2\lfloor \varphi/2 \rfloor}$. Moreover, S can be found in linear time.*

Recall that *biconnected* means that the graph is connected, and remains connected even after the deletion of any single vertex. If a graph is connected but not biconnected, then it has an *articulation point*, that is, a vertex whose removal disconnects the graph into more than one component. The *maximal face size* of a planar graph

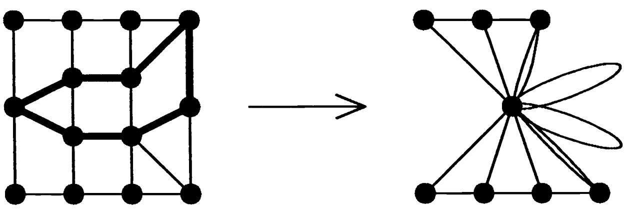


FIG. 1. *Contracting a cycle to a single node.*

refers to the maximum number of edges around any face. Finally, the term *simple cycle separator* refers to a set of vertices S that form a separator in the sense of the previous definition, and that are the vertex set of a simple cycle in G.

For the remainder of this section we restrict attention to biconnected planar graphs with face size bounded by φ. We remark that a strengthening of the foregoing theorem appears in Gazit and Miller [4], which allows G to have a constant number of faces with, say, $\sqrt{n}$ vertices without disturbing the bound.

The assumption that G is biconnected is not actually a restriction. First, if G is disconnected, then each component can be treated separately. Second, if G is connected but not biconnected, then the biconnected components can be identified in linear time (see Aho, Hopcroft, and Ullman [1]). It suffices to work with biconnected components, since any particular cycle (null vector) is contained in a unique biconnected component of G.

Our main goal for this section is to analyze a procedure **T** that finds a spanning tree. A basic building block for **T** is Algorithm **P** that decomposes the graph using the previous theorem. The first two steps of Algorithm **P** are as follows:

P-1. Find in G an (α,β)-SCS, say S.

P-2. Transform G into G' by contracting S into a single vertex σ.

These steps are illustrated in Fig. 1.

LEMMA 2.2. *Vertex σ is the only articulation point of G'.*

Proof. Suppose $\tau \neq \sigma$ is an articulation point (a.p.) in G'. Let $G'_1, G'_2, \ldots, G'_k$ be the biconnected components of G' articulated by σ. Since σ is an a.p. in G', τ belongs to exactly one G'_i, say G'_1. Let $G'_{1,1}$, $G'_{1,2}$ be two of the biconnected components of G'_1 articulated by τ. Since τ is an a.p., σ belongs to exactly one $G'_{1,i}$, say $G'_{1,1}$.

Now $G'_{1,2}$ is a subgraph of G, because it has not been affected by the contraction of S. But then τ is an articulation point in G between $G'_{1,2}$ and S, which contradicts the hypothesis that G is biconnected. □

By virtue of Lemma 2.2 we can define the final step of procedure **P**.

P-3. Separate the biconnected components of G' into disconnected subgraphs,

$$G^1_1, G^1_2, \ldots, G^1_k,$$

with every subgraph receiving its own copy of σ, the a.p. in G'. The final graph, G^1, is the disjoint union of the subgraphs G^1_i (see Fig. 2).

LEMMA 2.3. *The maximal face size of G^1_i, a connected subgraph of G^1, does not exceed the maximal face size of the original graph G.*

Proof. Observe that, given an embedding of G, there is an embedding of G^1_i that corresponds to topologically contracting the embedded edges of S. Focus on this embedding of G^1_i.

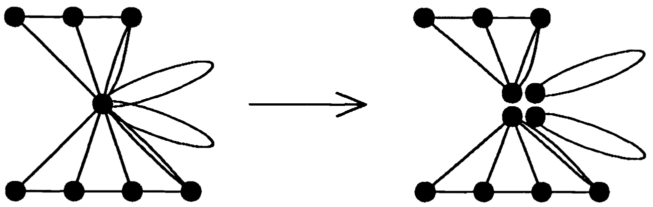


FIG. 2. *Graph G_1 based on Fig. 1.*

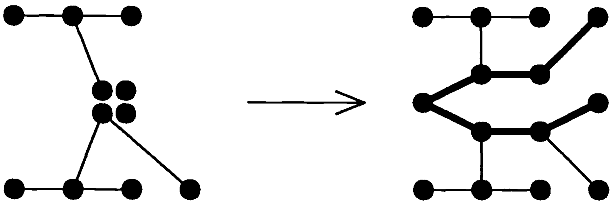


FIG. 3. *Finding a tree recursively.*

Let $v_1, v_2, \ldots, v_l$ be the vertices in a face of G_i^1. If $v_k \neq \sigma$ for $k = 1, \ldots, l$, then the face was not changed by the contraction of S, and the lemma follows. Now say $v_l = \sigma$. The list of vertices $v_1, \ldots, v_l$ in the embedding of G_i^1 have no edges adjacent to them in the interior of the face. Since no edge adjacent to these vertices was deleted by the contraction process, this means that vertices $v_1, \ldots, v_{l-1}$ form a path in G such that there are no edges adjacent to these vertices that are embedded on one "side" (either left or right) of the path. Thus this path is part of a face in G of size at least l. □

By virtue of Lemma 2.3 we can recursively apply procedure **P** to each of the connected subgraphs in G^1. That will be the basis for procedure **T** to construct the spanning tree. The parameter κ below is a constant depending on φ that we will specify later.

```
If |V(G)| ≤ κ
    T-1: Return an arbitrary spanning tree $T$ in $G$.
Else
    T-2: Apply procedure P to $G$ (using the SCS $S$).
         In each subgraph $G_i^1$, recursively use procedure T to get a tree $T_i$.
    T-3: Construct in $G$ the spanning tree $T$, joining:
         (a) Forests $\hat{T}_i$, the uncontracted versions of $T_i$, and
         (b) The SCS $S$ with an arbitrary edge deleted.
```

This procedure is illustrated in Fig. 3 for the graph used in Figs. 1 and 2.

LEMMA 2.4. *The graph returned by procedure* **T** *is a spanning tree.*

Proof. We will prove the theorem by induction on the number of levels we recursively called procedure **T**. At the last recursive call, **T-1** returns a spanning tree. This is the basis for our inductive argument. Now assume that the recursive calls to **T** return spanning trees T_i at each subgraph G_i^1. To complete our induction we must prove that T is a spanning tree for G.

Let C be cycle S with an arbitrary edge deleted, so that C is a simple path spanning the vertices of S. We argue by contradiction that T cannot have cycles. Suppose T has a cycle. Since each $\hat{T}_i$ is a forest and the various $\hat{T}_i$'s have no common vertices, the cycle must be formed by edges of C and a particular $\hat{T}_i$. But this is not possible because then the same cycle could be contracted to form a cycle in T_i.

Also, we claim that T is connected. This is because the forests $\hat{T}_i$ and C span all the vertices, and each tree in the forest made up of $\hat{T}_i$'s contains a vertex of C. □

We have now demonstrated that procedure **T** constructs a spanning tree. We now put an upper bound on the length of the cycle formed by adding any edge of $G-T$ to T. We use the term *co-tree edge* to refer to such edges, and the term *co-tree cycle* to refer to the unique cycle induced by a co-tree edge. The co-tree cycles correspond to the null vectors in the basis, and the number of edges in the co-tree cycles correspond to the number of nonzero entries of the nullspace basis.

THEOREM 2.5. *A co-tree cycle of T has length at most $\delta\sqrt{n}+\kappa$, where*

$$\delta = \beta/\left(1-\sqrt{\alpha+1/\kappa}\right).$$

Proof. Let $f(n)$ be the maximum number of edges in a co-tree cycle. We will prove the theorem by induction on the number of levels we recursively called procedure **T**.

If $n \le \kappa$, the theorem is trivial.

Otherwise, let S denote the SCS constructed at the top level of **T**, and observe that a co-tree cycle must lie interior to one of the subgraphs G_i^1, plus S. Now we can use Theorem 2.1 to get the recursion

$$f(n) \le f(\alpha n+1)+\beta\sqrt{n}.$$

The first term accounts for the edges in G_i^1, a graph that has at most αn vertices plus the copy of σ, and the second term accounts for the edges in S. Our induction hypothesis states that

$$f(\alpha n+1) \le \delta\sqrt{\alpha n+1}+\kappa.$$

So to prove our theorem it suffices to establish, for all $\kappa \le m \le n$, that

$$\delta\sqrt{\alpha m+1}+\beta\sqrt{m} \le \delta\sqrt{m},$$

i.e.,

$$\delta \ge \beta/\left(1-\sqrt{\alpha+1/m}\right).$$

Since $\alpha = \frac{2}{3} < 1$, we can choose $\kappa = 30$ to ensure that the last denominator is positive, and given κ and β we can choose δ to satisfy the inequality for $m=\kappa$. For example, if the graph is triangulated, then Theorem 2.1 gives us $\beta = \sqrt{8}$, and taking $\kappa = 30$ we can take $\delta = 20$. □

As a corollary to Theorem 2.5, we can easily put a bound on $g(n)$, the total length of the all co-tree cycles, i.e., the number of nonzeros in the nullspace basis. Theorem 2.5 gives us a bound for $f(n)$, the length of the longest co-tree cycle. A planar graph with n vertices has at most $3n-6$ edges [7]. Thus there are at most $(3n-6)-(n-1)$ co-tree cycles, so the total co-tree length $g(n)$ is bounded by $(2n-5)f(n)$, i.e., $O(n\sqrt{n})$.

Finally, let T be the tree in G constructed by algorithm **T**. Now we can see that the incidence matrix of the cycles in the cycle basis can be written in RBAF. In

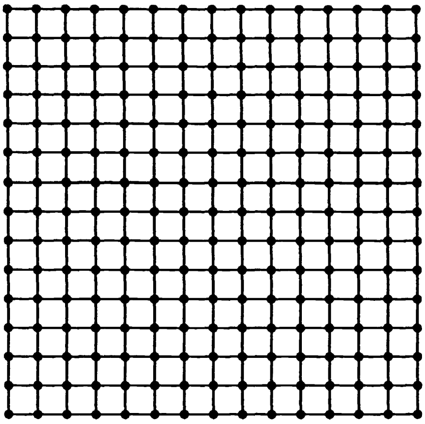


FIG. 4. *The grid graph in the $k = 16$ case.*

particular, we number the edges of G according to the biconnected components of G', with the edges of S numbered last. Then each cycle in the nullspace basis has nonzeros in rows of V corresponding to edges of one particular biconnected component, plus edges from S. We can also nest the RBAF structure in V according to the recursive levels of procedure **T**.

3. The grid graph. For the $(k-1)\times(k-1)$ grid graph $G(k)$, with $n = (k-1)^2$ vertices and $m = (k-1)(k-2)$ edges, we can define a spanning tree $T(k)$ that gives us a better bound on the total length of the co-tree cycle basis.

For this section we consider only the case that $k = 2^j$, although our results can be generalized. The $k = 16$ graph is illustrated in Fig. 4.

Tree $T(k)$ is constructed recursively. In the case that $k = 2$, i.e., a 1×1 grid graph, the spanning tree is the unique node of the graph. This is the basis of the recursion. In order to construct $T(k)$ for $k > 2$, we need the following basic building blocks:

$\alpha(k)$: The central vertex of $G(k)$, i.e., the vertex at position $(k/2, k/2)$.

$\beta(k)$: The vertex of $G(k)$ at position $(k/2, k/4)$.

$\gamma(k)$: The vertex of $G(k)$ at position $(k/2, 3k/4)$.

$X(k)$: The edges of $G(k)$ of the central row and column.

$H(k)$: The edges in $X(k)$ plus the edges incident to $\beta(k)$ or $\gamma(k)$.

Graph $H(k)$ is illustrated in Fig. 5; vertices $\beta(k)$ and $\gamma(k)$ are shown as squares, and vertex $\alpha(k)$ is shown as a bullseye.

Notice that $G(k) - X(k)$ has four connected components, each of which is isomorphic to $G(k/2)$. We can recursively define $T(k)$ as $H(k)$, plus a copy of $T(k/2)$ for each component. We omit the proof that $T(k)$ is indeed a spanning tree of $G(k)$. An example of $T(k)$ in the case $k = 16$ is illustrated in Fig. 6. This tree is similar to a graph that has occurred in the very large scale integration (VLSI) literature known as the H-tree [16].

We consider the tree $T(k)$ to be rooted at $\alpha(k)$, and define $d^k(\omega)$ as the distance

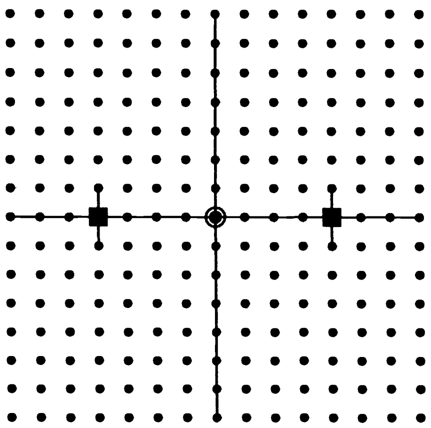


FIG. 5. *The graph $H(k)$ with $\alpha(k), \beta(k), \gamma(k)$ illustrated.*

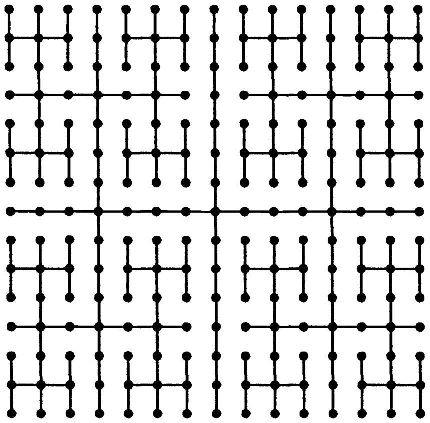


FIG. 6. *The graph $T(k)$ in the case $k = 16$.*

of a vertex ω in $T(k)$ to the root, i.e., $d(\omega, \alpha(k))$. We denote by $r(k)$ the radius of $T(k)$, i.e., the maximal distance $d^k(\omega)$.

LEMMA 3.1. *The radius $r(k)$ of $T(k)$ is bounded by k.*

Proof. Recall $k = 2^j$. We will prove the lemma by induction on j. If $j = 1$, then $T(2) = H(2) = \alpha$ and $r(2) = 0$. This is the basis for our induction.

For $k > 1$ let us consider $d^k(\omega)$ for an arbitrary vertex $\omega \in T(k)$. If $\omega \in X(k)$ then $d^k(\omega) \leq k/2$, because each of the four branches of $X(k)$ has diameter $k/2$. If $\omega \notin X(k)$, then ω belongs to one of the four copies of $T(k/2)$ rooted at $\alpha(k/2)$, and connected to $H(k)$ through $\beta(k)$ or $\gamma(k)$. Let us say that it is in the upper left copy of $T(k/2)$, so that it is rooted through $\beta(k)$. Let $\alpha(k/2)$ denote the root of the particular copy of $T(k/2)$ under consideration. Then we have the chain of equations

$$\begin{aligned} d^k(\omega) &= d(\omega, \alpha(k)) \\ &= d(\omega, \alpha(k/2)) + d(\alpha(k/2), \alpha(k)) \\ &= d^{k/2}(\omega) + d(\alpha(k/2), \beta(k)) + d(\beta(k), \alpha(k)) \\ &= d^{k/2}(\omega) + k/4 + k/4 \\ &\leq r(k/2) + k/2. \end{aligned}$$

Hence $r(k) \leq r(k/2) + k/2$, and that gives the induction step for our hypothesis, $r(k) \leq k$. □

THEOREM 3.2. *The total length $g(k)$ of the co-tree cycle basis for the tree $T(k)$ in the $(k-1) \times (k-1)$ square grid $G(k)$ is asymptotically bounded by $6k^2 \log k$.*

Proof. Let us first look at the co-tree cycles that have edges in $X(k)$. We observe that the only co-tree cycles with edges in $X(k)$ are those generated by co-tree edges incident to vertices of $X(k)$. Also, there are only $4k$ of those edges. Let $f(k)$ be length of a given co-tree cycle, S, intersecting $X(k)$. Observe that the edges of S not in $X(k)$ must lie in one of the four copies of $T(k/2)$. The number of such edges, as in the previous lemma, is at most $r(k/2) + k/4$. The number of edges of S in $X(k)$ is at most $3k/4$, because they must be in the border of one of the quadrant regions, and cannot cross $\beta(k)$. Therefore, the total number of edges in S is bounded as follows:

$$f(k) \leq (r(k/2) + k/4) + 3k/4 \leq 3k/2.$$

Now, all the remaining co-tree cycles lie in one of the copies of $T(k/2)$, and we can write

$$g(k) \leq 4g(k/2) + (4k)f(k) \leq 4g(k/2) + 6k^2.$$

But a recursion in the form $g(k) \leq 4g(k/2) + ck^2$ is known to be bounded by $ck^2 \log(k) + O(k^2)$. Indeed, George's original [5] derivation of nested dissection on grids came up with the same expression for the fill during Gaussian elimination. So we can limit the total length of the $T(k)$ co-tree cycle basis for $G(k)$ by

$$g(k) \leq 6k^2 \log k + O(k^2) = 6n \log n + O(n). \qquad \square$$

4. Generalizing the construction. In this section we discuss how to generalize the results of §2 to matrices other than adjacency matrices of planar graphs. We first note that the planarity assumption is necessary only to obtain simple cycle separators. We see that the necessary property of the separator is not that it be a simple cycle, but rather that the subgraph induced by the separator nodes be connected. Indeed, the separators in §3 are not cycles. With this weaker hypothesis all of the results of §2 go through. Recent results by Miller, Teng, and Vavasis [10] suggest that very broad classes of graphs could have connected separators.

If we want to generalize beyond adjacency matrices of graphs, we see that the first crucial property in our analysis was that matrix A can be written in the form of (2). This is not enough to carry out our algorithm; this form is analogous to a graph G having a node separator but not necessarily a connected node separator. The generalization of the "connected" property is that matrix S_{k+1} have full row rank.

If this happens, then it can be shown using linear algebra that it suffices to find a nullspace basis for S_{k+1} and for $\hat{B}_1, \ldots, \hat{B}_k$, where $\hat{B}_i$ denotes the matrix

$$\begin{pmatrix} B_i \\ w_i^T \end{pmatrix}.$$

In this formulation w_i^T is a row vector with a "don't care" entry for each column in which S_i has at least one nonzero entry. From fundamental nullspace bases for $\hat{B}_i$ and for S_i we can assemble a fundamental nullspace basis for A. This construction works only under the usual noncancellation assumption that the nullspace basis of A can be predicted entirely from the positions of the nonzero entries in A. See, for example, Pothen [13].

Note added in proof. After this paper was written and accepted, we learned of recent work [2] that improves on some of the results in this paper.

REFERENCES

[1] A. Aho, J. Hopcroft, and J. Ullman, *Data Structures and Algorithms*, Addison-Wesley, Reading, MA, 1983.

[2] N. Alon, R. M. Karp, D. Peleg, and D. West, *A graph-theoretic game and its application to the k-server problem,* Tech. Rep. 91-066, International Computer Science Institute, Berkeley, CA, December 1991.

[3] T. Coleman and A. Pothen, *The null space problem* II. *Algorithms*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 544–563.

[4] H. Gazit and G. Miller, *Planar separators and the Euclidean norm*, preprint, 1988.

[5] A. George, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 345–363.

[6] J. Gilbert and M. Heath, *Computing a sparse basis for the null space*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 446–459.

[7] F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA, 1972.

[8] I. Kaneko, M. Lawo, and G. Thierauf, *On computational procedures for the force method,* Internat. J. Numer. Methods Engrg., 18 (1982), pp. 1469–1495.

[9] G. Miller, *Finding small simple cycle separators for* 2*-connected planar graphs*, J. Comput. System Sci., 32 (1986), pp. 265–279.

[10] G. Miller, S.-H. Teng, and S. Vavasis, *A unified geometric approach to graph separators*, preprint.

[11] R. Plemmons and R. White, *Substructuring methods for computing the nullspace of equilibrium matrices*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 1–22.

[12] A. Pothen, *Sparse Null Basis Computations in Structural Optimization*, Tech. Rep. CS-88-27, Dept. of Computer Science, the Pennsylvania State Univ., University Park, PA, 1988.

[13] ———, *Sparse Null Bases and Marriage Theorems*, Ph.D. thesis, Cornell University, Ithaca, NY, 1984; Tech. Rep. 85-676, Computer Science Dept., Cornell University, Ithaca, NY, 1985.

[14] J. Stern, *Simulated Annealing with a Temperature Dependent Cost Function*, ORSA J. Comput., 4 (1992), pp. 311–319.

[15] G. Strang, *A framework for equilibrium equations*, SIAM Rev., 30 (1988), pp. 283–297.

[16] J. Ullman, *Computational Aspects of* VLSI, Computer Science Press, Rockville, MD, 1984.

[17] D. Welsh, *Matroid Theory*, London Math. Soc. Monograph 8, Academic Press, New York, 1976.

013

LINEAR QUADRATIC PROBLEMS WITH INDEFINITE COST FOR DISCRETE TIME SYSTEMS*

A. C. M. RAN† AND H. L. TRENTELMAN‡

Abstract. This paper deals with the discrete-time, infinite-horizon linear quadratic problem with indefinite cost criterion. Given a discrete-time linear system, an indefinite cost-functional and a linear subspace of the state space, the problem of minimizing the cost-functional over all inputs that force the state trajectory to converge to the given subspace is considered. A geometric characterization of the set of all Hermitian solutions of the discrete-time algebraic Riccati equation is given. This characterization forms the discrete-time counterpart of the well-known geometric characterization of the set of all real symmetric solutions of the continuous-time algebraic Riccati equation as developed by Willems [*IEEE Trans. Automat. Control*, 16 (1971), pp. 621–634] and Coppel [*Bull. Austral. Math. Soc.*, 10 (1974), pp. 377–401]. In the set of all Hermitian solutions of the Riccati equation the solution that leads to the optimal cost for the above-mentioned linear quadratic problem is identified. Finally, necessary and sufficient conditions for the existence of optimal controls are given.

Key words. linear quadratic optimal control, indefinite cost functional, discrete-time, Riccati equation

AMS subject classifications. 93C05, 93C35, 93C60

1. Introduction.

This paper has two main goals. First, we want to establish the discrete-time counterpart of the by now "classical" geometric characterization of the lattice of real symmetric solutions of the continuous-time algebraic Riccati equation as given in [1] and [8]. Subsequently, we want to apply these results to the discrete-time linear quadratic optimization problem with linear endpoint constraints. Given a discrete-time linear system, the latter problem consists of minimizing a general *indefinite* quadratic cost-functional over the class of input functions that force the state trajectory to converge to an a priori given subspace of the state space (or, equivalently, that force a given linear function of the state to converge to zero). A complete treatment of this optimization problem for the continuous-time case was given only very recently in [5] and [6].

With respect to our first goal, it will be shown that, as in the continuous-time case, if the algebraic Riccati equation has at least one Hermitian solution, then it has a smallest one and a largest one. Furthermore, any Hermitian solution of the algebraic Riccati equation can be written as a "linear combination" of these extremal solutions. In order to derive these results we will make use of the characterization of all Hermitian solutions of the discrete-time Riccati equation in terms of certain invariant Lagrangian subspaces, as established in [4] (see also [2]).

With respect to our second goal, we want to note that compared to the usual discrete-time linear quadratic optimal control problems, our problem formulation introduces generalizations into two independent directions. First, in contrast to the existing literature on this subject, we do not require the quadratic form in the cost-functional to be positive semidefinite (the "linear quadratic regulator problem"). Instead, the quadratic form is allowed to be indefinite. Second, our problem formula-

* Received by the editors December 19, 1990; accepted for publication (in revised form) November 22, 1991.

† Faculty of Mathematics and Computing Science, Vrije Universiteit, De Boelelaan 1081, 1081 HV Amsterdam, the Netherlands (ran@cs.vu.nl).

‡ Mathematics Institute, P.O. Box 800, 9700 AV Groningen, the Netherlands (trentelman@rug.nl).

tion includes a fixed, but arbitrary, linear endpoint constraint, in the sense that the optimization is performed over the class of all input functions that force the state trajectory to converge to an a priori given subspace. A solution to the usual zero endpoint problem (in which the optimal state trajectory is required to converge to the origin) can thus be obtained from our results by setting this subspace to be equal to the zero subspace. On the other hand, a solution to the free endpoint problem (no constraint on the optimal state trajectory) can be obtained from our results by taking the subspace to be equal to the entire state space.

The outline of this paper is as follows. In §2 we shall formulate the optimization problem that we want to consider. This section also contains a statement of the main result of this paper, that is, a characterization of the optimal cost, necessary and sufficient conditions for the existence of optimal controls, and an expression for the optimal state feedback control law. In §3 we shall establish the characterization of the set of all Hermitian solutions of the discrete-time algebraic Riccati equation as announced above. Finally, in §4 we shall give a proof of the main result as stated in §2.

2. Problem statement and main results. In this paper, we will consider the discrete-time system

$$x_{k+1} = Ax_k + Bu_k, \tag{2.1}$$

where the state variable x_k takes its values in $\mathcal{C}^n$ and the input variable u_k takes its values in $\mathcal{C}^m$. In (2.1) we have $A \in \mathcal{C}^{n\times n}$ and $B \in \mathcal{C}^{n\times m}$. As a standing assumption, we take (A, B) to be a controllable pair. We will consider optimization problems of the type

$$\inf \sum_{k=0}^{\infty} \begin{pmatrix} x_k \\ u_k \end{pmatrix}^* \begin{pmatrix} Q & C^* \\ C & R \end{pmatrix} \begin{pmatrix} x_k \\ u_k \end{pmatrix}. \tag{2.2}$$

Here, R, Q, and C are complex matrices of appropriate dimensions, and $R = R^*$, $Q = Q^*$. In this paper, a standing assumption will be that R is nonsingular. The expression (2.2) of course needs some explanation. For any $x_0 \in \mathcal{C}^n$ and any control sequence $u = \{u_k\}_{k=0}^{\infty}$, we define

$$J_T(x_0, u) := \sum_{k=0}^{T} \begin{pmatrix} x_k \\ u_k \end{pmatrix}^* \begin{pmatrix} Q & C^* \\ C & R \end{pmatrix} \begin{pmatrix} x_k \\ u_k \end{pmatrix}. \tag{2.3}$$

Let

$$U(x_0) := \left\{ u \mid \lim_{T\to\infty} J_T(x_0, u) \text{ exists in } \mathcal{R} \cup \{-\infty, +\infty\} \right\},$$

and for any control sequence $u \in U(x_0)$, define the associated cost by

$$J(x_0, u) := \lim_{T\to\infty} J_T(x_0, u). \tag{2.4}$$

The optimization problem of minimizing the cost functional (2.4) over the class of inputs $U(x_0)$ is called the *free endpoint linear quadratic problem.* The optimal cost associated with this problem is equal to

$$V_f(x_0) := \inf_{u\in U(x_0)} J(x_0, u). \tag{2.5}$$

Compare this problem with the usual *zero endpoint problem*, where instead of $U(x_0)$, the cost-functional is minimized over the class of all inputs that force the corresponding state trajectory to converge to the origin, i.e., over

$$U_s(x_0) := \left\{ u \in U(x_0) | \lim_{k \to \infty} x(x_0, u)_k = 0 \right\}. \tag{2.6}$$

The associated optimal cost is given by

$$V_+(x_0) := \inf_{u \in U_s(x_0)} J(x_0, u).$$

In the present paper, we will study a generalization of the above two linear quadratic problems, the *linear quadratic problem with linear endpoint constraints.* Given a linear subspace $\mathcal{L}$ of the state space $\mathcal{C}^n$, the latter problem consists of minimizing the cost functional (2.4) over all inputs u that force the state trajectory to converge to the subspace $\mathcal{L}$:

$$U_{\mathcal{L}}(x_0) := \left\{ u \in U(x_0) | \lim_{k \to \infty} d(x(x_0, u)_k, \mathcal{L}) = 0 \right\}. \tag{2.7}$$

In the above, for a given point $x \in \mathcal{C}^n$, $d(x, \mathcal{L})$ denotes the usual distance from the point x to the subspace $\mathcal{L}$. The optimal cost for the latter problem is given by

$$V_{\mathcal{L}}(x_0) := \inf_{u \in U_{\mathcal{L}}(x_0)} J(x_0, u). \tag{2.8}$$

Obviously, both the free endpoint problem and the fixed endpoint problem are special cases of the latter problem formulation: take $\mathcal{L} = \mathcal{C}^n$ and $\mathcal{L} = 0$, respectively.

An important role will be played by the set of Hermitian solutions of the discrete-time algebraic Riccati equation

$$P = A^*PA + Q - (C + B^*PA)^*(R + B^*PB)^{-1}(C + B^*PA). \tag{2.9}$$

P is called a solution of the algebraic Riccati equation if $R + B^*PB$ is nonsingular and if P satisfies (2.9). Besides controllability of (A, B) and nonsingularity of R, we shall assume throughout that $A - BR^{-1}C$ is nonsingular and that $\Psi(\eta) > 0$ for some $\eta \in \mathcal{T}$, where

$$\Psi(z) := \begin{pmatrix} B^*(Iz^{-1} - A^*)^{-1} & I \end{pmatrix} \begin{pmatrix} Q & C^* \\ C & R \end{pmatrix} \begin{pmatrix} (Iz - A)^{-1}B \\ I \end{pmatrix}.$$

Here, $\mathcal{T}$ denotes the unit circle. Let us comment on these assumptions. The nonsingularity of R and $A - BR^{-1}C$ is mainly assumed for technical reasons. These assumptions allow us to give a description of the set of Hermitian solutions of (2.9) in terms of invariant subspaces of a certain matrix (see §3) rather than in terms of a pencil of matrices (compare, e.g., [12]). Note that, in contrast to most literature on the continuous-time case, we do not assume $R > 0$, but only R nonsingular. (If in the continuous-time case the condition $R \geq 0$ is violated, then we have $V_+(x_0) = -\infty$ for every initial state x_0. In the discrete-time case the condition $R \geq 0$ is *not* necessary for boundedness from below of the cost-functional $J(x_0, u)$.) The positivity of $\Psi(\eta)$ for some $\eta \in \mathcal{T}$ should be contrasted with the continuous-time case. The corresponding matrix-valued function for the continuous-time case is given by

$$\tilde{\Psi}(z) := \begin{pmatrix} B^*(Iz + A^*)^{-1} & I \end{pmatrix} \begin{pmatrix} Q & C^* \\ C & R \end{pmatrix} \begin{pmatrix} (Iz - A)^{-1}B \\ I \end{pmatrix}.$$

Taking the limit as z goes to infinity, this function is seen to have value R at infinity. So, under the usual assumption of the continuous-time case, $R > 0$, a similar assumption is automatically satisfied: there is an η on the imaginary axis such that $\tilde{\Psi}(\eta) > 0$.

Under the assumptions outlined in the previous paragraph, the set of Hermitian solutions of (2.9) will turn out to have a maximal element P_+ and a minimal element P_- (see (3.5) below). (For the existence of both a maximal and a minimal Hermitian solution, the controllability of (A, B) is a natural sufficient condition. Assuming only *stabilizability* of (A, B) would only guarantee the existence of a maximal solution.) Put $\Delta := P_+ - P_-$. Let us denote by A_+ and A_- the matrices

$$A_+ := A - B(R + B^*P_+B)^{-1}(C + B^*P_+A),$$

$$A_- := A - B(R + B^*P_-B)^{-1}(C + B^*P_-A).$$

It will be seen that $\sigma(A_+) \subset \bar{\mathcal{D}}$ and $\sigma(A_-) \subset \bar{\mathcal{D}}^e$ (here $\mathcal{D}$ denotes the open unit disk, $\mathcal{D}^e$ denotes the exterior of the closed unit disc, and for any matrix A, $\sigma(A)$ denotes the set of eigenvalues of A). Given a subspace $\mathcal{L}$ of $\mathcal{C}^n$ we introduce the subspace

$$\mathcal{V}(\mathcal{L}) := \langle \mathcal{L} \cap \ker P_- \mid A_- \rangle \cap \mathcal{X}_+(A_-). \tag{2.10}$$

(Here, for a matrix A, $\mathcal{X}_+(A)$ denotes the spectral subspace of A corresponding to its eigenvalues in $\mathcal{D}^e$; likewise, we shall denote $\mathcal{X}_0(A)$, $\mathcal{X}_-(A)$ the spectral subspaces of A corresponding to its eigenvalues in $\mathcal{T}$, respectively, $\mathcal{D}$.) As usual, the notation $\langle \mathcal{W} \mid A \rangle$, where $\mathcal{W}$ is a subspace of $\mathcal{C}^n$ and A an $n \times n$ matrix, denotes the largest A-invariant subspace in $\mathcal{W}$. If H is a matrix such that $\ker H = \mathcal{L} \cap \ker P_-$, then $\mathcal{V}(\mathcal{L})$ is the undetectable subspace of the pair (H, A_-) with respect to the stability set $\bar{\mathcal{D}}$. Now, let

$$P_{\mathcal{L}} := P_- \pi_{\mathcal{V}(\mathcal{L})} + P_+(I - \pi_{\mathcal{V}(\mathcal{L})}), \tag{2.11}$$

where $\pi_{\mathcal{V}(\mathcal{L})}$ is the projection onto $\mathcal{V}(\mathcal{L})$ along $(\Delta \mathcal{V}(\mathcal{L}))^\perp$. It will turn out that $P_{\mathcal{L}}$ is a solution of (2.9). Finally, if $\mathcal{L}$ is a subspace of $\mathcal{C}^n$ and if P is a Hermitian $n \times n$ matrix, then we will say that P is *negative semidefinite on* $\mathcal{L}$ if the following conditions hold:

- for all $x_0 \in \mathcal{L} : x_0^* P x_0 \leq 0$,
- for all $x_0 \in \mathcal{L} : x_0^* P x_0 = 0 \Leftrightarrow P x_0 = 0$.

According to this definition, a Hermitian matrix P is negative semidefinite on $\mathcal{C}^n$ if and only if it is negative semidefinite in the usual sense. Furthermore, any Hermitian matrix is negative semidefinite on the zero subspace $\{0\}$. The main result of this paper can now be formulated as follows.

THEOREM 2.1. *Suppose (A, B) is controllable, R is nonsingular, $A - BR^{-1}C$ is nonsingular, and $\Psi(\eta) > 0$ for some $\eta \in \mathcal{T}$. Assume further that* (2.9) *has at least one Hermitian solution and assume P_- is negative semidefinite on $\mathcal{L}$. Then we have*

(i) *$V_{\mathcal{L}}(x_0)$ is finite for all x_0, and $V_{\mathcal{L}}(x_0) = x_0^* P_{\mathcal{L}} x_0$,*

(ii) *for all x_0 there is an input u^+ such that $V_{\mathcal{L}}(x_0) = J(x_0, u^+)$ if and only if $\ker \Delta \subseteq \mathcal{L} \cap \ker P_-$; in that case u^+ is unique and is given by the state feedback control law*

$$u_k^+ = -(R + B^* P_{\mathcal{L}} B)^{-1}(C + B^* P_{\mathcal{L}} A) x_k.$$

This result is the discrete-time analogue of [6, Thm. 4.1]. We stress that the above theorem also provides information on the free endpoint problem and on the zero endpoint problem. Indeed, for the free endpoint problem we set $\mathcal{L} = \mathcal{C}^n$. The corresponding subspace $\mathcal{V}(\mathcal{L})$ is then equal to $\mathcal{V} = \langle \ker P_- \mid A_- \rangle \cap \mathcal{X}_+(A_-)$. Define

$$P_f := P_- \pi_{\mathcal{V}} + P_+(I - \pi_{\mathcal{V}}),$$

where, again, $\pi_{\mathcal{V}}$ is the projection onto $\mathcal{V}$ along $(\Delta\mathcal{V})^\perp$. P_f is a solution of (2.9) and we find the following corollary.

COROLLARY 2.2. *Suppose (A, B) is controllable, R is nonsingular, $A - BR^{-1}C$ is nonsingular, and $\Psi(\eta) > 0$ for some $\eta \in \mathcal{T}$. Assume further that* (2.9) *has at least one Hermitian solution and assume that $P_- \leq 0$. Then we have*

(i) *$V_f(x_0)$ is finite for all x_0, and $V_f(x_0) = x_0^* P_f x_0$,*

(ii) *for all x_0 there is an input u^+ such that $V_f(x_0) = J(x_0, u^+)$ if and only if* $\ker \Delta \subseteq \ker P_-$; *in that case u^+ is unique and is given by the state feedback control law*

$$u_k^+ = -(R + B^* P_f B)^{-1}(C + B^* P_f A)x_k.$$

The above corollary is the discrete-time analogue of [5, Thm. 5.1]. In order to get the corresponding result on the zero endpoint problem, we set $\mathcal{L} = \{0\}$. The corresponding subspace $\mathcal{V}(\mathcal{L})$ is then equal to $\mathcal{V} = \{0\}$ and we find that the relevant solution of (2.9) is equal to P_+. Thus we find the following discrete-time version of [8, Thm. 7].

COROLLARY 2.3. *Suppose (A, B) is controllable, R is nonsingular, $A - BR^{-1}C$ is nonsingular, and $\Psi(\eta) > 0$ for some $\eta \in \mathcal{T}$. Assume further that* (2.9) *has at least one Hermitian solution. Then we have*

(i) *$V_+(x_0)$ is finite for all x_0, and $V_+(x_0) = x_0^* P_+ x_0$,*

(ii) *for all x_0 there is an input u_+ such that $V_+(x_0) = J(x_0, u_+)$ if and only if* $\Delta > 0$; *in that case u_+ is unique and is given by the state feedback control law*

$$u_k^+ = -(R + B^* P_+ B)^{-1}(C + B^* P_+ A)x_k.$$

In this paper, we shall give a proof of Theorem 2.1. The proof that we shall give basically follows the line of [6]; the details will be provided in §4. In §3 we give a description of all solutions of (2.9) in terms of P_- and P_+. The continuous-time analogue of this description is due to Coppel [1]. The argument here is somewhat more complicated and uses ideas from [2] and [4].

3. Description of solutions of the algebraic Riccati equation. Consider the discrete-time algebraic Riccati equation (2.9). In addition to the controllability of (A, B) and the assumption that R is nonsingular, we assume throughout that $A - BR^{-1}C$ is nonsingular and $\Psi(\eta) > 0$ for some η on the unit circle. We are then in a position to apply [4, Thm. 4.1] (see also [2, Thm. 4.4]), which gives a description of solutions of the algebraic Riccati equation in terms of certain invariant subspaces. To be precise, put

(3.1)

$$T = \begin{pmatrix} A - BR^{-1}C + BR^{-1}B^*(A - BR^{-1}C)^{*-1}(Q - C^*R^{-1}C) & -BR^{-1}B^*(A - BR^{-1}C)^{*-1} \\ -(A - BR^{-1}C)^{*-1}(Q - C^*R^{-1}C) & (A - BR^{-1}C)^{*-1} \end{pmatrix}.$$

The next theorem provides necessary and sufficient conditions for the existence of a Hermitian solution of (2.9), and two different descriptions of the set of all Hermitian solutions. Both descriptions are in terms of certain invariant subspaces of T.

THEOREM 3.1. *Assume (A, B) is controllable, R is nonsingular, $A - BR^{-1}C$ is nonsingular, and $\Psi(\eta) > 0$ for some $\eta \in \mathcal{T}$. Then the following statements are equivalent:*

(i) *There exists a Hermitian solution of* (2.9),

(ii) *T has an invariant subspace $\mathcal{M}$ such that*

$$\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \mathcal{M} = \mathcal{M}^{\perp}, \tag{3.2}$$

(iii) *the partial multiplicities of T (i.e., the sizes of the Jordan blocks in the Jordan form of T) corresponding to its eigenvalues on the unit circle $\mathcal{T}$ are all even,*

(iv) *$\Psi(z) \geq 0$ for all z on the unit circle $\mathcal{T}$.*

DESCRIPTION 1. *Assume that one of the above equivalent conditions hold. Then any T-invariant subspace $\mathcal{M}$ for which* (3.2) *holds is of the form*

$$\mathcal{M} = \operatorname{im} \begin{pmatrix} I \\ P \end{pmatrix} \tag{3.3}$$

for some Hermitian solution P of (2.9), *and, conversely, if $P = P^*$ solves* (2.9), *then $\mathcal{M}$ given by* (3.3) *is T-invariant and satisfies* (3.2).

DESCRIPTION 2. *Assume that one of the above equivalent conditions hold. Then for every T-invariant subspace $\mathcal{N}$ with the property that $\sigma(T \mid \mathcal{N}) \subset \mathcal{D}^e$ there is a unique solution $P = P^*$ of* (2.9) *with*

$$\operatorname{im} \begin{pmatrix} I \\ P \end{pmatrix} \cap \mathcal{X}_+(T) = \mathcal{N}. \tag{3.4}$$

Conversely, for every Hermitian solution P of (2.9) *the subspace $\mathcal{N}$ given by* (3.4) *is T-invariant and has the property that $\sigma(T \mid \mathcal{N}) \subset \mathcal{D}^e$. Here $\mathcal{X}_+(T)$ denotes the sum of the generalized eigenspaces of T with respect to its eigenvalues in $\mathcal{D}^e$.*

Now let P be any Hermitian solution of (2.9). Then it is a straightforward calculation to see that

$$\Psi(z) = \Xi(\bar{z}^{-1})^*(R + B^*PB)\Xi(z),$$

where

$$\Xi(z) = I + (R + B^*PB)^{-1}(C + B^*PA)(Iz - A)^{-1}B$$

(see, e.g., [2]). Consequently, for any Hermitian solution P of (2.9) we have $R + B^*PB \geq 0$. Combined with the fact that $R + B^*PB$ is also nonsingular, we find that if (i)–(iv) in Theorem 3.1 hold then we have $R + B^*PB > 0$ for *any* solution $P = P^*$ of (2.9) (see also [2, Thm. 2.5]).

Let P_+ and P_- be the unique solutions for which

$$\operatorname{im} \begin{pmatrix} I \\ P_+ \end{pmatrix} \cap \mathcal{X}_+(T) = \{0\}, \qquad \operatorname{im} \begin{pmatrix} I \\ P_- \end{pmatrix} \cap \mathcal{X}_+(T) = \mathcal{X}_+(T), \tag{3.5}$$

respectively. We shall show that P_+ is the maximal solution and P_- the minimal solution of the equation (2.9). First we prove a lemma.

LEMMA 3.1. *Let P_- be the solution defined by* (3.5) *and suppose P is an arbitrary Hermitian solution. Introduce*

$$\begin{aligned} A_- &= A - B(R + B^*P_-B)^{-1}(C + B^*P_-A), \\ S_- &= R + B^*P_-B. \end{aligned}$$

Then $X := P - P_-$ satisfies the algebraic Riccati equation

$$X = A_-^*XA_- - A_-^*XB(S_- + B^*XB)^{-1}B^*XA_-. \tag{3.6}$$

Conversely, any Hermitian solution X of (3.6) *gives a solution of* (2.9) *via $P = X + P_-$.*

Proof. Introduce the following matrices:

$$\begin{aligned} S_- &:= R + B^*P_-B, & S &:= R + B^*PB, \\ E_- &:= C + B^*P_-A, & E &:= C + B^*PA, \\ L_- &:= S_-^{-1}E_-, & L &:= S^{-1}E. \end{aligned}$$

To prove (3.6), compute

$$\begin{aligned} X - A_-^*XA_- &= X - (A - BL_-)^*X(A - BL_-) \\ &= (P - P_-) - A^*(P - P_-)A + A^*(P - P_-)BL_- \\ &\quad + L_-^*B^*(P - P_-)A - L_-^*B^*(P - P_-)BL_-. \end{aligned} \tag{3.7}$$

Since P and P_- solve (2.9), we have

$$X - A^*XA = E_-^*S_-^{-1}E_- - E^*S^{-1}E. \tag{3.8}$$

Furthermore,

$$\begin{aligned} A^*(P - P_-)BL_- &= (A^*PB - A^*P_-B)L_- \\ &= (E^* - E_-^*)L_- \\ &= (E^* - E_-^*)S_-^{-1}E_- \end{aligned}$$

and

$$L_-^*B^*(P - P_-)BL_- = E_-^*S_-^{-1}B^*(P - P_-)BS_-^{-1}E_-.$$

Using these equalities and (3.8) in (3.7), we obtain

$$\begin{aligned} X - A_-^*XA_- &= E_-^*S_-^{-1}E_- - E^*S^{-1}E + (E^* - E_-^*)S_-^{-1}E_- \\ &\quad + E_-^*S_-^{-1}(E - E_-) - E_-^*S_-^{-1}B^*(P - P_-)BS_-^{-1}E_- \\ &= -E_-^*S_-^{-1}E_- - E^*S^{-1}E + E^*S_-^{-1}E_- + E_-^*S_-^{-1}E \\ &\quad - E_-^*S_-^{-1}B^*(P - P_-)BS_-^{-1}E_- \\ &= -E_-^*S_-^{-1}(S_- + B^*(P - P_-)B)S_-^{-1}E_- - E^*S^{-1}E \\ &\quad + E^*S_-^{-1}E_- + E_-^*S_-^{-1}E \\ &= -E_-^*S_-^{-1}SS_-^{-1}E_- - E^*S^{-1}E + E^*S_-^{-1}E_- + E_-^*S_-^{-1}E \\ &= -(E^* - E_-^*S_-^{-1}S)S^{-1}(E - SS_-^{-1}E_-). \end{aligned} \tag{3.9}$$

Here we used the fact that $S_- + B^*XB = S$. Moreover,

$$SS_-^{-1} = (R + B^*PB)(R + B^*P_-B)^{-1} = I + B^*XBS_-^{-1},$$

so

$$\begin{aligned} E - SS_-^{-1}E_- &= E - E_- - B^*XBS_-^{-1}E_- \\ &= B^*XA - B^*XBS_-^{-1}E_- \\ &= B^*X(A - BS_-^{-1}E_-) \\ &= B^*XA_-. \end{aligned}$$

Hence, from (3.9), we see

$$X - A_-^*XA_- = -A_-^*XB(S_- + B^*XB)^{-1}B^*XA_-,$$

which proves (3.6). The converse follows by a similar computation. □

With a similar argument one shows that the following lemma is true.

LEMMA 3.2. *Let P_+ be the solution introduced above; suppose P is an arbitrary Hermitian solution. Introduce*

$$\begin{aligned} A_+ &= A - B(R + B^*P_+B)^{-1}(C + B^*P_+A), \\ S_+ &= R + B^*P_+B. \end{aligned}$$

Then $X = P - P_+$ satisfies the algebraic Riccati equation

$$X = A_+^*XA_+ - A_+^*XB(S_+ + B^*XB)^{-1}B^*XA_+. \tag{3.10}$$

Conversely, any solution of (3.10) *gives a solution of* (2.9) *via $P = X + P_+$.*

According to Theorem 3.1, the subspaces

$$\operatorname{im}\begin{pmatrix} I \\ P_+ \end{pmatrix} \quad \text{and} \quad \operatorname{im}\begin{pmatrix} I \\ P_- \end{pmatrix}$$

are T-invariant. The following lemma states that the restrictions of T to these subspaces are semistable and semi-antistable, respectively.

LEMMA 3.3.

$$\sigma\left(T \mid \operatorname{im}\begin{pmatrix} I \\ P_+ \end{pmatrix}\right) \subset \bar{\mathcal{D}}, \qquad \sigma\left(T \mid \operatorname{im}\begin{pmatrix} I \\ P_- \end{pmatrix}\right) \subset \bar{\mathcal{D}}^e.$$

Proof. The first statement is an immediate consequence of (3.5). To prove the second statement, define

$$J := \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}.$$

Note that T is J-unitary, i.e., $T^*JT = J$. This implies that T is invertible and that $T^{-1} = -JT^*J$. Now assume that

$$x \in \operatorname{im}\begin{pmatrix} I \\ P_- \end{pmatrix}, \quad Tx = \lambda x \quad \text{and} \quad \lambda \in \mathcal{D}.$$

Then $T^{-1}x = \lambda^{-1}x$, from which $T^*Jx = \lambda^{-1}Jx$. This implies that $Jx \in \mathcal{X}_+(T^*)$, since $\lambda^{-1} \in \mathcal{D}^e$. We also have

$$Jx \in J \operatorname{im}\begin{pmatrix} I \\ P_- \end{pmatrix} = \left(\operatorname{im}\begin{pmatrix} I \\ P_- \end{pmatrix}\right)^{\perp}.$$

Using (3.5) the latter implies $Jx \in \mathcal{X}_+(T)^{\perp}$. Since $\mathcal{X}_+(T)^{\perp} = \mathcal{X}_0(T^*) \oplus \mathcal{X}_-(T^*)$ we obtain $Jx = 0$, so $x = 0$. □

It is a straightforward but tedious calculation to show that

$$T \mid \operatorname{im}\begin{pmatrix} I \\ P_+ \end{pmatrix}, \qquad T \mid \operatorname{im}\begin{pmatrix} I \\ P_- \end{pmatrix}$$

are similar to A_+ and A_-, respectively (see, e.g., [10] and [11]). Consequently, $\sigma(A_+) \subset \bar{\mathcal{D}}$ and $\sigma(A_-) \subset \bar{\mathcal{D}}^e$. Introduce

$$(3.11) \qquad A_P = A - B(R + B^*PB)^{-1}(C + B^*PA).$$

The following theorem states that P_- is the smallest Hermitian solution of (2.9). Likewise, P_+ is the largest Hermitian solution of (2.9). Furthermore, $P = P_-$ is the only Hermitian solution with the property that $\sigma(A_P) \subset \bar{\mathcal{D}}^e$ and $P = P_+$ is the only Hermitian solution with the property that $\sigma(A_P) \subset \bar{\mathcal{D}}$.

THEOREM 3.2. *Assume* (A, B) *is controllable,* R *is nonsingular,* $A - BR^{-1}C$ *is nonsingular, and* $\Psi(\eta) > 0$ *for some* $\eta \in \mathcal{T}$. *Assume that* (2.9) *has at least one Hermitian solution. Let* P_- *and* P_+ *be the solutions determined by* (3.5). *Then* $\sigma(A_+) \subset \bar{\mathcal{D}}$ *and* $\sigma(A_-) \subset \bar{\mathcal{D}}^e$. *Furthermore, for any Hermitian solution of* (2.9), *we have*

$$P_- \leq P \leq P_+.$$

In addition, if P *is a Hermitian solution with the property that* $\sigma(A_P) \subset \bar{\mathcal{D}}^e$, *then* $P = P_-$. *If* P *is a Hermitian solution with the property that* $\sigma(A_P) \subset \bar{\mathcal{D}}$, *then* $P = P_+$.

Proof. Let P be a Hermitian solution. Then $X = P - P_-$ solves (3.6) and $S_- + B^*XB = R + B^*PB > 0$. We shall prove that $X \geq 0$. First we shall prove that $\mathcal{X}_0(A_-) \subseteq \ker X$. In order to prove this, choose a basis of $\mathcal{X}_0(A_-)$ consisting of eigenvectors and generalized eigenvectors. Such a basis consists of chains of vectors $x_1, \ldots, x_k$ with the property that

$$\begin{aligned} A_-x_1 &= \lambda x_1, \\ A_-x_2 &= \lambda x_2 + x_1, \\ &\vdots \\ A_-x_k &= \lambda x_k + x_{k-1}, \end{aligned}$$

where $|\lambda| = 1$. We will show by induction that $Xx_1 = \cdots = Xx_k = 0$. Assume $A_-x_1 = \lambda x_1$. Using (3.6) we obtain

$$x_1^*Xx_1 = |\lambda|^2 x_1^*Xx_1 - |\lambda|^2 x_1^*XB(S_- + B^*XB)^{-1}B^*Xx_1.$$

This yields $x_1^*XB(S_- + B^*XB)^{-1}B^*Xx_1 = 0$, from which we obtain $x_1^*XB = 0$. Again using (3.6), this implies $x_1^*X = \bar{\lambda}x_1^*XA_-$, from which

$$x_1^*X(A_- - \bar{\lambda}^{-1}I \quad B) = 0.$$

By controllability of (A, B), the latter implies that $Xx_1 = 0$. Now, assume that $A_-x_r = x_{r-1} + \lambda x_r$ and $Xx_{r-1} = 0$. Using (3.6) we find

$$x_r^*Xx_r = (x_{r-1}^* + \bar{\lambda}x_r^*)X(x_{r-1} + \lambda x_r) \\ - (x_{r-1}^* + \bar{\lambda}x_r^*)XB(S_- + B^*XB)^{-1}B^*X(x_{r-1} + \lambda x_r).$$

The latter implies that $(x_{r-1}^* + \bar{\lambda}x_r^*)XB = 0$ and hence $x_r^*XB = 0$. Also, $x_r^*X = \bar{\lambda}x_r^*XA_-$. Again, by controllability of (A, B), this yields $Xx_r = 0$. This proves our claim that $\mathcal{X}_0(A_-) \subseteq \ker X$.

To proceed, let U be a unitary matrix such that

$$U^*A_-U = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

with $\sigma(A_{11}) \subset \mathcal{T}$ and $\sigma(A_{22}) \subset \mathcal{D}^e$. Let $Y = U^*XU$. Obviously, since $\mathcal{X}_0(A_-) \subseteq \ker X$, we have

$$Y = \begin{pmatrix} 0 & 0 \\ 0 & Y_{22} \end{pmatrix}.$$

Furthermore, it follows from (3.6) that $Y_{22} - A_{22}^*Y_{22}A_{22} \leq 0$. Since the eigenvalues of A_{22} lie *strictly* outside the unit disc, the latter can be shown to imply $Y_{22} \geq 0$ (see, e.g., [9, Thm. 13.2.3]). Thus we have proven $X \geq 0$.

A completely similar argument can be used to show that any Hermitian solution P satisfies $P \leq P_+$. Finally, note that in the above proof we only used the facts that $\sigma(A_-) \subset \bar{\mathcal{D}}^e$ and $\sigma(A_+) \subset \bar{\mathcal{D}}$. Hence, if P is a Hermitian solution with the property that $\sigma(A_P) \subset \bar{\mathcal{D}}^e$, then we must have $P \leq Q$ for any Hermitian solution Q, in particular, for $Q = P_-$. This shows that $P = P_-$. The statement on P_+ is proven similarly. □

Next we prove an analogue for the discrete-time case of a theorem first proved by Coppel [1] for the continuous-time case.

THEOREM 3.3. *Assume that (A, B) is controllable, R is nonsingular, $A - BR^{-1}C$ is nonsingular, and $\Psi(\eta) > 0$ for some η on the unit circle. Let P_- and P_+ be the minimal and maximal Hermitian solutions of* (2.9), *respectively. Put $\Delta := P_+ - P_-$ and*

$$A_- := A - B(R + B^*P_-B)^{-1}(C + B^*P_-A).$$

Then, for every A_--invariant subspace $\mathcal{V}$ of $\mathcal{X}_+(A_-)$, we have

$$\mathcal{C}^n = \mathcal{V} \oplus (\Delta\mathcal{V})^\perp. \tag{3.12}$$

Let $\pi_{\mathcal{V}}$ denote the projection onto $\mathcal{V}$ along $(\Delta\mathcal{V})^\perp$. For every A_--invariant subspace $\mathcal{V} \subseteq \mathcal{X}_+(A_-)$ the matrix P defined by

$$P = P_-\pi_{\mathcal{V}} + P_+(I - \pi_{\mathcal{V}}) \tag{3.13}$$

is a Hermitian solution of (2.9). *Conversely, for every Hermitian solution P of* (2.9) *there exists a unique A_--invariant subspace $\mathcal{V}$ of $\mathcal{X}_+(A_-)$ such that* (3.13) *holds. This subspace $\mathcal{V}$ is equal to $\mathcal{V} = \mathcal{X}_+(A_P)$, where A_P is defined by*

$$A_P = A - B(R + B^*PB)^{-1}(C + B^*PA).$$

In addition, $\mathcal{X}_0(A_P) = \ker \Delta$ *and* $\mathcal{X}_-(A_P) = (\Delta \mathcal{V})^\perp \cap \mathcal{X}_-(A_+)$.

It will become clear in the proof that this result is actually little more than a reformulation of the last part of Theorem 3.1.

Proof. First we show that it suffices to prove the theorem for equation (3.6). Note that $X_- = 0$ is the minimal solution of (3.6) and $X_+ = P_+ - P_-$ is the maximal solution. Moreover, by straightforward computation,

$$(3.14) \qquad \begin{aligned} (A_-)_X &:= A_- - B(S_- + B^*XB)^{-1}B^*XA_- \\ &= A_{X+P_-}. \end{aligned}$$

In particular, this means that $(A_-)_{X_-} = A_-$ and $(A_-)_{X_+} = A_+$, i.e., the A_- and A_+ matrices remain the same. Because of Lemma 3.1, P is a solution of (2.9) if and only if $P = X + P_-$ for some solution of (3.6). Now assume the theorem is true for (3.6). Let $\mathcal{V}$ be an A_--invariant subspace of $\mathcal{X}_+(A_-)$. As the A_- matrix remains the same, we conclude that (3.12) holds. Furthermore, the matrix $X := X_+(I - \pi_\mathcal{V})$ is a Hermitian solution of (3.6). This implies that $P := P_- + X = P_-\pi_\mathcal{V} + P_+(I - \pi_\mathcal{V})$ is a Hermitian solution of (2.9). Conversely, let P be a solution of (2.9). Define $X := P - P_-$. There exists an A_--invariant subspace of $\mathcal{X}_+(A_-)$ such that $X = X_+(I - \pi_\mathcal{V})$. This, however, yields $P = P_-\pi_\mathcal{V} - P_+(I - \pi_\mathcal{V})$. Finally, since $(A_-)_X = A_P$ and $(A_-)_{X_+} = A_+$, we also find $\mathcal{V} = \mathcal{X}_+(A_P)$, $\ker \Delta = \mathcal{X}_0(A_P)$, and $\mathcal{X}_-(A_+) \cap (\Delta\mathcal{V})^\perp = \mathcal{X}_-(A_P)$.

It remains to prove the statements of the theorem for equation (3.6). The matrix T given by (3.1) looks particularly simple in the case of equation (3.6):

$$T = \begin{pmatrix} A_- & -BS_-^{-1}B^*A_-^{*-1} \\ 0 & A_-^{*-1} \end{pmatrix},$$

with $S_- > 0$. Note that A_- has all its eigenvalues on or outside the unit circle. Hence we have $\mathcal{X}_+(T) \subseteq \operatorname{im} \binom{I}{0}$; more precisely,

$$\mathcal{X}_+(T) = \mathcal{X}_+(A_-) \times \{0\}.$$

So, T-invariant subspaces $\mathcal{N}$ with $\sigma(T \mid \mathcal{N}) \subset \mathcal{D}^e$ are precisely the subspaces of the form

$$\mathcal{N} = \mathcal{V} \times \{0\},$$

where $\mathcal{V}$ is A_--invariant and $\mathcal{V} \subseteq \mathcal{X}_+(A_-)$.

Now, let $\mathcal{V}$ be an A_--invariant subspace of $\mathcal{X}_+(A_-)$. We will prove (3.12) and the fact that $X_+(I - \pi_\mathcal{V})$ is a Hermitian solution of (3.6). According to Theorem 3.1, there is a unique solution X of (3.6) such that

$$\operatorname{im} \begin{pmatrix} I \\ X \end{pmatrix} \cap (\mathcal{X}_+(A_-) \times \{0\}) = \mathcal{V} \times \{0\}.$$

From [3, §§2, 7] and Theorem 3.1 we have

$$\operatorname{im} \begin{pmatrix} I \\ X \end{pmatrix} = (\mathcal{V} \times \{0\}) \oplus \left(\operatorname{im} \begin{pmatrix} I \\ X \end{pmatrix} \cap \mathcal{X}_0(T) \right) \oplus \left(\mathcal{X}_-(T) \cap \left(\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \mathcal{N} \right)^\perp \right),$$

and moreover $\operatorname{im} \binom{I}{X} \cap \mathcal{X}_0(T)$ is the same subspace for any Hermitian solution X (here we also use the fact that the signs in the sign characteristic of

$$\left(T, \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \right)$$

are all the same; see [4, Thm. 1.2]). Then

$$\begin{aligned}\operatorname{im}\begin{pmatrix} I \\ X \end{pmatrix} \cap \mathcal{X}_0(T) &= \operatorname{im}\begin{pmatrix} I \\ X_+ \end{pmatrix} \cap \mathcal{X}_0(T)\\ &= \operatorname{im}\begin{pmatrix} I \\ X_- \end{pmatrix} \cap \mathcal{X}_0(T)\\ &= \operatorname{im}\begin{pmatrix} I \\ 0 \end{pmatrix} \cap \mathcal{X}_0(T)\\ &\subseteq \ker X_+ \times \{0\}.\end{aligned}$$

Conversely, for $x \in \ker X_+$, we have from (3.6)

$$0 = A^*{}_-X_+(A_- - B(S_- + B^*X_+B)^{-1}B^*X_+A_-)x = A^*{}_-X_+A_+x = 0.$$

So $X_+A_+x = 0$, i.e., $\ker X_+$ is A_+-invariant. Consider the equality

$$T\begin{pmatrix} I \\ X_+ \end{pmatrix} x = \begin{pmatrix} I \\ X_+ \end{pmatrix} A_+x,$$

which holds by Theorem 3.1, in particular, the first description of the set of solutions. For $x \in \ker X_+$ this gives

$$T\begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} A_+x \\ 0 \end{pmatrix}.$$

However,

$$T\begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} A_-x \\ 0 \end{pmatrix}.$$

It follows that for $x \in \ker X_+$ we have $A_+x = A_-x$, and hence $\ker X_+$ is also A_--invariant. We claim that

$$\ker X_+ \subseteq \mathcal{X}_0(A_-). \tag{3.15}$$

Indeed, assume $x \in \ker X_+$, $x \neq 0$, and $A_-x = \lambda x$. Then $T\binom{x}{0} = \lambda\binom{x}{0}$. Since

$$\begin{pmatrix} x \\ 0 \end{pmatrix} \in \operatorname{im}\begin{pmatrix} I \\ X_+ \end{pmatrix},$$

by Lemma 3.3 we have $\lambda \in \bar{\mathcal{D}}$. On the other hand, since $\lambda \in \sigma(A_-)$, $\lambda \in \bar{\mathcal{D}}^e$. Thus $\lambda \in \mathcal{T}$, which proves the claim. It follows from (3.15) that

$$\operatorname{im}\begin{pmatrix} I \\ X \end{pmatrix} \cap \mathcal{X}_0(T) = \ker X_+ \times \{0\}.$$

Next,

$$\mathcal{X}_-(T) = \mathcal{X}_-(T) \cap \operatorname{im}\begin{pmatrix} I \\ X_+ \end{pmatrix},$$

so

$$\mathcal{X}_-(T) \cap \left(\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\mathcal{N}\right)^{\perp} \subseteq \operatorname{im}\begin{pmatrix} I \\ X_+ \end{pmatrix} \cap (\mathcal{C}^n \times \mathcal{V}^{\perp}).$$

Now

$$\begin{pmatrix} x \\ X_+x \end{pmatrix} \in \mathcal{C}^n \times \mathcal{V}^\perp$$

implies $X_+x \in \mathcal{V}^\perp$, i.e., $x \in (X_+\mathcal{V})^\perp$. So

$$\mathcal{X}_-(T) \cap \left(\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \mathcal{N} \right)^\perp \subseteq \left\{ \begin{pmatrix} x \\ X_+x \end{pmatrix} \mid x \in (X_+\mathcal{V})^\perp \right\}.$$

Since obviously ker $X_+ \subseteq (X_+\mathcal{V})^\perp$ also, we find

$$\text{(3.16)} \qquad \text{im} \begin{pmatrix} I \\ X \end{pmatrix} \subseteq (\mathcal{V} \times \{0\}) + \left\{ \begin{pmatrix} x \\ X_+x \end{pmatrix} \mid x \in (X_+\mathcal{V})^\perp \right\}.$$

Using the previous inclusion, it is easy to see that $\mathcal{C}^n = \mathcal{V} + (X_+\mathcal{V})^\perp$. We claim that the latter is, in fact, a direct sum. Indeed, $x \in \mathcal{V} \cap (X_+\mathcal{V})^\perp$ implies $(x, X_+x) = 0$, from which $X_+x = 0$. Thus $\mathcal{V} \cap (X_+\mathcal{V})^\perp \subseteq \ker X_+ \cap \mathcal{V} = 0$ (recall that $\mathcal{V} \subseteq \mathcal{X}_+(A_-)$ while $\ker X_+ \subseteq \mathcal{X}_0(A_-)$). Now, if $\pi_\mathcal{V}$ is the projection onto $\mathcal{V}$ along $(X_+\mathcal{V})^\perp$, it can be seen that (3.16) implies

$$X = X_+(I - \pi_\mathcal{V}).$$

Next, conversely, let X be a Hermitian solution of (3.10). Define

$$\mathcal{N} := \text{im} \begin{pmatrix} I \\ X \end{pmatrix} \cap \mathcal{X}_+(T).$$

Then $\mathcal{N}$ is a T-invariant subspace and $\sigma(T \mid \mathcal{N}) \subset \mathcal{D}^e$. Thus $\mathcal{N}$ has the form $\mathcal{N} = \mathcal{V} \times \{0\}$ for some A_--invariant subspace $\mathcal{V}$ of $\mathcal{X}_+(A_-)$. By repeating the argument in the first part of this proof, it is then shown that we must have $X = X_+(I - \pi_\mathcal{V})$.

Finally, we will show that if $\mathcal{V} \subset \mathcal{X}_+(A_-)$ is A_--invariant and $X = X_+(I - \pi_\mathcal{V})$, then $\mathcal{V} = \mathcal{X}_+((A_-)_X)$, $\ker X_+ = \mathcal{X}_0((A_-)_X)$, and $(\Delta\mathcal{V})^\perp \cap \mathcal{X}_-(A_+) = \mathcal{X}_-((A_-)_X)$.

As in the proof of Theorem 3.2, we show that $\mathcal{X}_0(A_-) \subseteq \ker X_+$. Combined with (3.15) this yields $\ker X_+ = \mathcal{X}_0(A_-)$. Since $0 \leq X \leq X_+$, $\ker X_+ \subseteq \ker X$. Hence $(A_-)_X \mid \ker X_+ = A_- \mid \ker X_+$. This yields

$$\ker X_+ = \mathcal{X}_0((A_-)_X),$$

as desired. We will now show that

$$\text{(3.17)} \qquad \mathcal{V} \subseteq \mathcal{X}_+((A_-)_X).$$

Indeed, note that $\mathcal{V} \subseteq \ker X$. Hence $(A_-)_X \mid \mathcal{V} = A_- \mid \mathcal{V}$, which yields (3.17). Next, we show that

$$\text{(3.18)} \qquad (X_+\mathcal{V})^\perp \subseteq \mathcal{X}_0((A_-)_X) \oplus \mathcal{X}_-((A_-)_X).$$

In order to prove this, first note that $A_-\mathcal{V} = \mathcal{V}$, since $\mathcal{V}$ is A_--invariant and since A_- is invertible. Now, the Riccati equation (3.6) with $X = X_+$ can be written as $X_+ = A_-^* X_+ A_+$. We claim that $(X_+\mathcal{V})^\perp$ is A_+-invariant. Let $w \in (X_+\mathcal{V})^\perp$. Then for all $v \in \mathcal{V}$,

$$0 = v^* X_+ w = v^* A_-^* X_+ A_+ w.$$

Hence $A_+w \in (X_+A_-\mathcal{V})^\perp = (X_+\mathcal{V})^\perp$. Next, by straightforward calculation, we show that

$$(3.19) \qquad (A_-)_X = A_+ - B(R + B^*(X + P_-)B)^{-1}B^*(X - X_+)A_+.$$

Since $X \mid (X_+\mathcal{V})^\perp = X_+ \mid (X_+\mathcal{V})^\perp$ (3.19) yields $(A_-)_X \mid (X_+\mathcal{V})^\perp = A_+ \mid (X_+\mathcal{V})^\perp$. Since $\sigma(A_+) \subset \bar{\mathcal{D}}$, this implies $\sigma((A_-)_X \mid (X_+\mathcal{V})^\perp) \subset \bar{\mathcal{D}}$, which yields (3.18).

By combining (3.17), (3.18), and the fact that $\mathcal{V} \oplus (X_+\mathcal{V})^\perp = \mathcal{C}^n$, we find that the inclusions (3.17) and (3.18) are, in fact, equalities. Finally, since $(A_-)_X \mid (X_+\mathcal{V})^\perp = A_+ \mid (X_+\mathcal{V})^\perp$, we find $(X_+\mathcal{V})^\perp \cap \mathcal{X}_-(A_+) = \mathcal{X}_-((A_-)_X)$. □

4. A proof of Theorem 2.1. The proof is split up into several lemmas, which are all discrete-time counterparts of results in [6]. In this section, let $\mathcal{L}$ be an arbitratry but fixed subspace of $\mathcal{C}^n$. We will first study the finiteness of the optimal cost $V_\mathcal{L}(x_0)$. Note that our assumption that (A, B) is controllable is sufficient to guarantee that $V_\mathcal{L}(x_0) < +\infty$ for all x_0. In the sequel we shall establish a sufficient condition to guarantee $V_\mathcal{L}(x_0) > -\infty$ for all x_0. From §2 recall the definition of negative semidefiniteness on $\mathcal{L}$ of a given Hermitian matrix P. It turns out that if the smallest solution P_- of the Riccati equation (2.9) is negative semidefinite on $\mathcal{L}$, then the optimal cost is finite.

LEMMA 4.1. *Assume that (A, B) is controllable, R is nonsingular, $A - BR^{-1}C$ is nonsingular, and $\Psi(\eta) > 0$ for some $\eta \in \mathcal{T}$. Furthermore, assume that* (2.9) *has at least one Hermitian solution. Then we have: if P_- is negative semidefinite on $\mathcal{L}$, then $V_\mathcal{L}(x_0) \in \mathcal{R}$ for all $x_0 \in \mathcal{C}^n$.*

Our proof of Lemma 4.1 uses the following two lemmas.

LEMMA 4.2. *Let $\mathcal{L}$ be a subspace of $\mathcal{C}^n$ and let H be a matrix such that $\mathcal{L} = \ker H$. Let $P \in \mathcal{C}^{n\times n}$ be Hermitian. Then P is negative semidefinite on $\mathcal{L}$ if and only if there exists $\lambda \in \mathcal{R}$ such that $P - \lambda H^*H$ is negative semidefinite.*

For a proof of the above lemma, we refer to [6].

LEMMA 4.3. *For any $x_0 \in \mathcal{C}^n$, any sequence u and any Hermitian solution P of* (2.9) *we have*

$$J_T(x_0, u) = x_0^*Px_0 - x_{T+1}^*Px_{T+1} + \sum_{k=0}^{T} \|u_k + (R + B^*PB)^{-1}(C + B^*PA)x_k\|^2_{R+B^*PB}.$$

Here, $\|v\|^2_S := v^*Sv$.

Proving this lemma is just a matter of standard computation.

Proof of Lemma 4.1. Let $x_0 \in \mathcal{C}^n$. Since (A, B) is controllable, there is an input sequence $u \in U_\mathcal{L}(x_0)$ such that $J(x_0, u) < +\infty$ (in fact, one can steer from x_0 to the origin in finite time). Thus $V_\mathcal{L}(x_0) \in \mathcal{R} \cup \{-\infty\}$. Now let $u \in U_\mathcal{L}(x_0)$ be arbitrary. Let H be such that $\mathcal{L} = \ker H$ and let $\lambda \in \mathcal{R}$ be such that $P_- - \lambda H^*H$ is negative semidefinite. According to Lemma 4.3, for all T we have

$$J_T(x_0, u) = x_0^*P_-x_0 - x_{T+1}^*(P_- - \lambda H^*H)x_{T+1} - \lambda\|Hx_{T+1}\|^2 + \sum_{k=0}^{T} \|u_k + (R + B^*PB)^{-1}(C + B^*PA)x_k\|^2_{R+B^*PB}.$$

Thus, for all $T \geq 0$,

$$J_T(x_0, u) \geq x_0^*P_-x_0 - \lambda\|Hx_{T+1}\|^2.$$

Since x_T converges to $\mathcal{L}$ as $T \to \infty$, we have $Hx_{T+1} \to 0$. It follows that

$$J(x_0, u) = \lim_{T\to\infty} J_T(x_0, u) \geq x_0^* P_- x_0.$$

Since the latter holds for all $u \in U_{\mathcal{L}}(x_0)$ this proves our claim. □

The next few lemmas give some general properties of linear systems.

LEMMA 4.4. *Consider the system*

$$x_{k+1} = Ax_k + v_k, \qquad y_k = Cx_k,$$

and suppose that (C, A) *is observable. Then if* $\{v_k\}_{k=0}^{\infty} \in \ell_2$ *and* $\{y_k\}_{k=0}^{\infty} \in \ell_\infty$, *necessarily* $\{x_k\}_{k=0}^{\infty} \in \ell_\infty$.

Proof. Since (C, A) is observable, there exists a matrix L such that $\sigma(A + LC) \subset \mathcal{D}$. Obviously, $\{x_k\}_{k=0}^{\infty}$ satisfies the difference equation

$$x_{k+1} = (A + LC)x_k - Ly_k + v_k.$$

Using some straightforward estimates we see that $v \in \ell_2$ and $y \in \ell_\infty$ imply $x \in \ell_\infty$. □

In the following lemma, if $\mathcal{C}_g$ is a subset of $\mathcal{C}$, then $\mathcal{X}_g(A)$ will denote the spectral subspace of A associated with its eigenvalues in $\mathcal{C}_g$, i.e., the largest A-invariant subspace $\mathcal{V}$ with the property that $\sigma(A \mid \mathcal{V}) \subset \mathcal{C}_g$. Using the previous lemma, we can now prove the following.

LEMMA 4.5. *Consider the system*

$$x_{k+1} = Ax_k + v_k, \qquad y_k = Cx_k.$$

Assume that (C, A) *is detectable (relative to* $\mathcal{C}_g$*). Let the state space* $\mathcal{C}^n$ *be decomposed into* $\mathcal{C}^n = \mathcal{X}_1 \oplus \mathcal{X}_2$, *where* $\mathcal{X}_1$ *is* A*-invariant. In this decomposition, let*

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Assume that $\sigma(A \mid \mathcal{X}_1) \subset \mathcal{C}_g$ *and* $\sigma(A \mid \mathcal{C}^n/\mathcal{X}_1) \subset \mathcal{C}/\mathcal{C}_g$. *Then for every initial condition* x_0 *we have: if* $\{v_k\}_{k=0}^{\infty} \in \ell_2$ *and* $\{y_k\}_{k=0}^{\infty} \in \ell_\infty$, *then* $\{x_{2,k}\}_{k=0}^{\infty} \in \ell_\infty$.

Proof. We claim that $\mathcal{X}_1 = \mathcal{X}_g(A)$. Indeed, by assumption, we have $\mathcal{X}_1 \subseteq \mathcal{X}_g(A)$. Denote $\sigma_0 := \sigma(A \mid \mathcal{X}_g(A)/\mathcal{X}_1)$. Then we have $\sigma_0 \subset \mathcal{C}_g$. Also, $\sigma_0 \subset \sigma(A \mid \mathcal{C}^n/\mathcal{X}_1) \subset \mathcal{C}/\mathcal{C}_g$. This implies that $\sigma_0 = \emptyset$ or, equivalently, $\mathcal{X}_1 = \mathcal{X}_g(A)$. By the fact that (C, A) is detectable with respect to $\mathcal{C}_g$, we may now conclude that $\langle \ker C \mid A \rangle \subseteq \mathcal{X}_1$. Decompose $\mathcal{X}_1 = \mathcal{X}_{11} \oplus \mathcal{X}_{12}$, with $\mathcal{X}_{11} := \langle \ker C \mid A \rangle$ and $\mathcal{X}_{12}$ arbitrary. Accordingly, partition

$$x_1 = \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}.$$

We then have $\mathcal{C}^n = \mathcal{X}_{11} \oplus \mathcal{X}_{12} \oplus \mathcal{X}_2$ with $x = (x_{11}^T, x_{12}^T, x_2^T)^T$. In this decomposition let

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix}, \quad C = \begin{pmatrix} 0 & C_2 & C_3 \end{pmatrix}, \quad v = \begin{pmatrix} v_{11} \\ v_{12} \\ v_2 \end{pmatrix}.$$

Obviously, the system

$$\left(\begin{pmatrix} C_2 & C_3 \end{pmatrix}, \begin{pmatrix} A_{22} & A_{23} \\ 0 & A_{33} \end{pmatrix} \right)$$

is observable. Moreover,

$$\begin{pmatrix} x_{12,k+1} \\ x_{2,k+1} \end{pmatrix} = \begin{pmatrix} A_{22} & A_{23} \\ 0 & A_{33} \end{pmatrix} \begin{pmatrix} x_{12,k} \\ x_{2,k} \end{pmatrix} + \begin{pmatrix} v_{12,k} \\ v_{2,k} \end{pmatrix},$$

$$y_k = \begin{pmatrix} C_2 & C_3 \end{pmatrix} \begin{pmatrix} v_{12,k} \\ v_{2,k} \end{pmatrix}.$$

It now follows from the previous lemma that

$$\begin{pmatrix} x_{12} \\ x_2 \end{pmatrix} \in \ell_\infty,$$

which implies that $x_2 \in \ell_\infty$. □

The next lemma tells us that a semistable controllable system has the property that all initial states can be steered to the origin with arbitrary small controls (in ℓ_2-sense).

LEMMA 4.6. *Consider the controllable system*

$$x_{k+1} = Ax_k + Bu_k, \quad x_0 \text{ given.}$$

Assume that $\sigma(A) \subset \bar{\mathcal{D}}$. Then for all $\epsilon > 0$ there exists a $u \in \ell_2$ such that $\|u\|_2 < \epsilon$ and $x(x_0, u)_k \to 0$ as $k \to \infty$.

Proof. We will first show that it suffices to prove the statement of the lemma with $\bar{\mathcal{D}}$ replaced by $\mathcal{T}$. Indeed, we can always choose a basis in $\mathcal{C}^n$ such that A and B have the form

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

with $\sigma(A_1) \subset \mathcal{D}$ and $\sigma(A_2) \subset \mathcal{T}$. Consider the subsystem $x_{1,k+1} = A_1 x_{1,k} + Bu_k$, with $x_{1,0}$ given. Obviously, for any input sequence $u \in \ell_2$, we automatically have that $x_1 \in \ell_2$ and hence $x_{1,k} \to 0$ as $k \to \infty$.

Assume, therefore, that $\sigma(A) \subset \mathcal{T}$. For any $\delta > 0$, any initial state x_0 and any input sequence u, define the quadratic cost

$$J_\delta(x_0, u) = \sum_{k=0}^{\infty} \|u_k\|^2 + \delta^2 \|x(x_0, u)_k\|^2 \tag{4.1}$$

and consider the optimization problem

$$\inf_u J_\delta(x_0, u). \tag{4.2}$$

Note that (4.2) can be considered as a "standard" discrete-time linear quadratic problem of minimizing $\sum_{k=0}^{\infty} u_k^* R u_k + x_k^* D^* D x_k$, with $R = I$ and $D = \delta I$ (see, for example, [7]). Since (A, B) is controllable and $(\delta I, A)$ is observable, the infimum (4.2) is equal to $x_0^* P(\delta) x_0$, with $P(\delta)$ the unique positive semidefinite solution of the Riccati equation

$$P = A^* P A + \delta^2 I - A^* P B (I + B^* P B)^{-1} B^* P A.$$

(In fact, $P(\delta) > 0$.) Furthermore, for any x_0 there exists a unique optimal input u^+ which is given by the state feedback control law

$$u^+ = -(I + B^*P(\delta)B)^{-1}B^*P(\delta)Ax^+,$$

and the corresponding optimal closed loop matrix

$$A - B(I + B^*P(\delta)B)^{-1}B^*P(\delta)A$$

is stable, i.e., has all its eigenvalues in $\mathcal{D}$. Now, we shall analyze what happens if $\delta \downarrow 0$. We claim that $P(\delta) \downarrow 0$. Indeed, $P(\delta) \geq 0$ and $P(\delta)$ is monotonically decreasing as $\delta \downarrow 0$. Hence there exists a Hermitian matrix $\bar{P} \geq 0$ such that $P(\delta) \downarrow \bar{P}$. Clearly, $\bar{P}$ satisfies the Riccati equation

$$P = A^*PA - A^*PB(I + B^*PB)^{-1}B^*PA. \tag{4.3}$$

Now, by applying Theorem 3.2 we find that the latter Riccati equation has a largest Hermitian solution, say, P_+. We contend that $P_+ = 0$. Indeed, $P = 0$ is a solution of (4.3) and it has the property that $\sigma(A_P) \subset \bar{\mathcal{D}}$ (since $A_P = A$ and $\sigma(A) \subset \mathcal{T}$). Thus we must have $\bar{P} \leq 0$. Our conclusion is that $\bar{P} = 0$.

In order to complete the proof, let $\epsilon > 0$. Choose $\delta > 0$ such that $x_0^*P(\delta)x_0 < \epsilon$. Choose the input sequence given by $u^+ = -(I + B^*P(\delta)B)B^*P(\delta)Ax^+$. Then we have

$$\|u^+\|^2 \leq J_\delta(x_0, u^+) = x_0^*P(\delta)x_0 < \epsilon.$$

Since the corresponding closed loop matrix is stable, the corresponding state trajectory converges to zero as $k \to \infty$. □

We proceed by decomposing $\mathcal{C}^n$ as follows. Let

$$\mathcal{X}_1 := \langle \mathcal{L} \cap \ker P_- \mid A_- \rangle \cap \mathcal{X}_+(A_-) = \mathcal{V}(\mathcal{L}),$$

and let $P_\mathcal{L}$ be the solution corresponding to $\mathcal{V}(\mathcal{L})$ according to Theorem 3.3. Put

$$A_\mathcal{L} := A - B(R + B^*P_\mathcal{L}B)^{-1}(C + B^*P_\mathcal{L}A).$$

According to Theorem 3.3, we have $\mathcal{X}_+(A_\mathcal{L}) = \mathcal{X}_1$. In addition, we define

$$\mathcal{X}_2 := \mathcal{X}_0(A_\mathcal{L}) = \mathcal{X}_0(A_-) = \ker \Delta,$$
$$\mathcal{X}_3 := \mathcal{X}_-(A_\mathcal{L}) = (\Delta\mathcal{V}(\mathcal{L}))^\perp \cap \mathcal{X}_-(A_+).$$

With respect to the decomposition $\mathcal{C}^n = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \mathcal{X}_3$, we have

$$A_- = \begin{pmatrix} A_{11} & 0 & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix}, \qquad A_\mathcal{L} = \begin{pmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & 0 & A'_{33} \end{pmatrix}.$$

Here, we have used that $A_- \mid \mathcal{X}_1 = A_\mathcal{L} \mid \mathcal{X}_1$ and that $A_- \mid \mathcal{X}_2 = A_\mathcal{L} \mid \mathcal{X}_2$. This follows most easily by combining (3.14) and the facts that $P_\mathcal{L} \mid \mathcal{V}(\mathcal{L}) = P_- \mid \mathcal{V}(\mathcal{L})$ and $P_\mathcal{L} \mid (\Delta\mathcal{V}(\mathcal{L}))^\perp = P_+ \mid (\Delta\mathcal{V}(\mathcal{L}))^\perp$. We also have

$$P_- = \begin{pmatrix} 0 & 0 & 0 \\ 0 & P_{22}^- & P_{23}^- \\ 0 & P_{23}^{-*} & P_{33}^- \end{pmatrix}, \qquad \Delta = \begin{pmatrix} \Delta_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Delta_{33} \end{pmatrix}$$

and

$$P_+ = P_- + \Delta = \begin{pmatrix} \Delta_{11} & 0 & 0 \\ 0 & P_{22}^- & P_{23}^- \\ 0 & P_{23}^{-*} & P_{33}^+ \end{pmatrix}, \qquad P_{\mathcal{L}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & P_{22}^- & P_{23}^- \\ 0 & P_{23}^{-*} & P_{33}^+ \end{pmatrix},$$

where $P_{33}^+ := P_{33}^- + \Delta_{33}$.

The next lemma states that $P_{\mathcal{L}}$ yields a lower bound for the linear quadratic problem (2.8) under consideration.

LEMMA 4.7. *Suppose (A, B) is controllable, R is nonsingular, $A - BR^{-1}C$ is nonsingular, and $\Psi(\eta) > 0$ for some $\eta \in \mathcal{T}$. Assume further that* (2.9) *has at least one Hermitian solution and assume that P_- is negative semidefinite on $\mathcal{L}$. Then for all x_0 and for all $u \in U_{\mathcal{L}}(x_0)$ we have*

$$J(x_0, u) \geq x_0^* P_{\mathcal{L}} x_0 + \sum_{k=0}^{\infty} \|u_k + (R + B^* P_{\mathcal{L}} B)^{-1}(C + B^* P_{\mathcal{L}} A)x_k\|_{R+B^*P_{\mathcal{L}}B}^2.$$

Proof. Let H be a matrix such that $L = \ker H$. Let $\lambda \in \mathcal{R}$ be such that $P_- - \lambda H^* H \leq 0$ (see Lemma 4.2). Take an arbitrary $u \in U_{\mathcal{L}}(x_0)$. It follows from Lemma 4.1 that $J(x_0, u) \in \mathcal{R} \cup \{+\infty\}$. If it is equal to $+\infty$, then the inequality trivially holds. Assume therefore that $J(x_0, u)$ is finite. Put

$$v_k = u_k + (R + B^* P_- B)^{-1}(C + B^* P_- A)x_k.$$

From Lemma 4.3 we have

(4.4)
$$\begin{aligned}\sum_{k=0}^{T} \|v_k\|_{R+B^*P_-B}^2 &= J_T(x_0, u) - x_0^* P_- x_0 + x_{T+1}^*(P_- - \lambda H^* H)x_{T+1} + \lambda \|Hx_{T+1}\|^2 \\ &\leq J_T(x_0, u) - x_0^* P_- x_0 + \lambda \|Hx_{T+1}\|^2.\end{aligned}$$

Since $J_T(x_0, u) \to J(x_0, u)$ and $Hx_T \to 0$, we find that $\{v_k\} \in \ell_2$. Again, using (4.4), this implies that

$$\lim_{T\to\infty} x_T^*(P_- - \lambda H^* H)x_T$$

exists and is finite. Thus $\lim_{T\to\infty} x_T^* P_- x_T$ exists and is finite. Also, since $P_- - \lambda H^* H$ is semidefinite, $(P_- - \lambda H^* H)x_k$ and hence $P_- x_k$ are bounded functions of k. Denote

$$y_k = \begin{pmatrix} P_- \\ H \end{pmatrix} x_k.$$

Then $\{y_k\} \in \ell_\infty$. Since $x_{k+1} = Ax_k + Bu_k$, we have that $\{x_k\}$, $\{y_k\}$, and $\{v_k\}$ are related by

$$x_{k+1} = A_- x_k + Bv_k, \qquad y_k = \begin{pmatrix} P_- \\ H \end{pmatrix} x_k.$$

Now decompose $\mathcal{C}^n$ as above: $\mathcal{C}^n = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \mathcal{X}_3$. Since

$$\mathcal{X}_1 \subseteq \ker \begin{pmatrix} P_- \\ H \end{pmatrix},$$

we have

$$\begin{pmatrix} P_- \\ H \end{pmatrix} = \begin{pmatrix} 0 & D_2 & D_3 \end{pmatrix} \tag{4.5}$$

for given matrices D_2 and D_3. With respect to the given decomposition, let

$$B = \begin{pmatrix} B_1 \\ B_2 \\ B_3 \end{pmatrix}.$$

Since $\mathcal{X}_1$ is the undetectable subspace (relative to $\bar{\mathcal{D}}$) of the system

$$\left(\begin{pmatrix} P_- \\ H \end{pmatrix}, A_- \right),$$

it is easily verified that the pair

$$\left(\begin{pmatrix} D_2 & D_3 \end{pmatrix}, \begin{pmatrix} A_{22} & A_{23} \\ 0 & A_{33} \end{pmatrix} \right)$$

is detectable (relative to $\bar{\mathcal{D}}$). Since $\sigma(A_-) \subset \bar{\mathcal{D}}^e$ and $\mathcal{X}_2 = \mathcal{X}_0(A_-)$, we have $\sigma(A_{22}) \subset \mathcal{T}$ and

$$\sigma\left(\begin{pmatrix} A_{11} & A_{13} \\ 0 & A_{33} \end{pmatrix} \right) \subset \mathcal{D}^e.$$

Hence $\sigma(A_{33}) \subset \mathcal{D}^e$. Also we have

$$\begin{pmatrix} x_{2,k+1} \\ x_{3,k+1} \end{pmatrix} = \begin{pmatrix} A_{22} & A_{23} \\ 0 & A_{33} \end{pmatrix} \begin{pmatrix} x_{2,k} \\ x_{3,k} \end{pmatrix} + \begin{pmatrix} B_2 \\ B_3 \end{pmatrix} v_k,$$
$$y_k = \begin{pmatrix} D_2 & D_3 \end{pmatrix} \begin{pmatrix} x_{2,k} \\ x_{3,k} \end{pmatrix}.$$

Since $\{v_k\} \in \ell_2$ and $\{y_k\} \in \ell_\infty$, by Lemma 4.5 (applied with $\mathcal{C}_g = \bar{\mathcal{D}}$), we have $\{x_{3,k}\} \in \ell_\infty$. Now consider Lemma 4.3. We have

$$J_T(x_0, u) = x_0^* P_{\mathcal{L}} x_0 - x_{T+1}^* P_{\mathcal{L}} x_{T+1} + \sum_{k=0}^{T} \|w_k\|^2_{R+B^*P_{\mathcal{L}}B},$$

where

$$w_k = u_k + (R + B^* P_{\mathcal{L}} B)^{-1}(C + B^* P_{\mathcal{L}} A) x_k.$$

Then

$$\begin{aligned} J_T(x_0, u) &= x_0^* P_{\mathcal{L}} x_0 + \sum_{k=0}^{T} \|w_k\|^2_{R+B^*P_{\mathcal{L}}B} \\ &\quad - \begin{pmatrix} x_{2,T+1} \\ x_{3,T+1} \end{pmatrix}^* \begin{pmatrix} P_{22}^- & P_{23}^- \\ P_{23}^{-*} & P_{33}^+ \end{pmatrix} \begin{pmatrix} x_{2,T+1} \\ x_{3,T+1} \end{pmatrix} \\ &= x_0^* P_{\mathcal{L}} x_0 + \sum_{k=0}^{T} \|w_k\|^2_{R+B^*P_{\mathcal{L}}B} \\ &\quad - x_{3,T+1}^* \Delta_{33} x_{3,T+1} - x_{T+1}^* P_- x_{T+1}. \end{aligned} \tag{4.6}$$

Now $x_{T+1}^*P_-x_{T+1}$, $J_T(x_0,u)$ are bounded as $T \to \infty$, and likewise $x_{3,T+1}$ is bounded as $T \to \infty$. It follows that

$$\sum_{k=0}^{\infty} \|w_k\|^2_{R+B^*P_{\mathcal{L}}B} < \infty,$$

and as $R + B^*P_{\mathcal{L}}B \geq R + B^*P_-B > 0$ we have $\{w_k\} \in \ell_2$. Now considering

$$x_{k+1} = Ax_k + Bu_k = A_{\mathcal{L}}x_k + Bw_k,$$

we see that $x_{3,k+1} = A'_{33}x_{3,k} + B_3w_k$. As $\sigma(A'_{33}) \subset \mathcal{D}$ (because of the fact that $\mathcal{X}_3 = \mathcal{X}_-(A_{\mathcal{L}})$) we obtain $\{x_{3,k}\} \in \ell_2$. Hence $\lim_{k\to\infty} x_{3,k} = 0$. Now from (4.6) we have

$$\begin{aligned} J_T(x_0,u) &= x_0^*P_{\mathcal{L}}x_0 - x_{T+1}^*(P_- - \lambda H^*H)x_{T+1} \\ &\quad - \lambda\|Hx_{T+1}\|^2 + \sum_{k=0}^{T} \|w_k\|^2_{R+B^*P_{\mathcal{L}}B} - x_{3,T+1}^*\Delta_{33}x_{3,T+1} \\ &\geq x_0^*P_{\mathcal{L}}x_0 - \lambda\|Hx_{T+1}\|^2 + \sum_{k=0}^{T} \|w_k\|^2_{R+B^*P_{\mathcal{L}}B} \\ &\quad - x_{3,T+1}^*\Delta_{33}x_{3,T+1}. \end{aligned}$$

The desired result then follows by taking the limit as $T \to \infty$ in the above inequality. □

Our next lemma states that $V_{\mathcal{L}}(x_0) = x_0^*P_{\mathcal{L}}x_0$, taking into account the previous lemma.

LEMMA 4.8. *Suppose (A,B) is controllable, R is nonsingular, $A - BR^{-1}C$ is nonsingular, and $\Psi(\eta) > 0$ for some $\eta \in \mathcal{T}$. Assume further that* (2.9) *has at least one Hermitian solution and assume that P_- is negative semidefinite on $\mathcal{L}$. Then for all x_0 and for all $\epsilon > 0$, there is a $u \in U_{\mathcal{L}}(x_0)$ such that $J(x_0,u) \leq x_0^*P_{\mathcal{L}}x_0 + \epsilon$.*

Proof. Put $w_k = u_k + (R + B^*P_{\mathcal{L}}B)^{-1}(C + B^*P_{\mathcal{L}}A)x_k$. From (4.6) we have for all $u \in U_{\mathcal{L}}(x_0)$

$$\begin{aligned} J_T(x_0,u) &= x_0^*P_{\mathcal{L}}x_0 + \sum_{k=0}^{T} \|w_k\|^2_{R+B^*P_{\mathcal{L}}B} \\ &\quad - \begin{pmatrix} x_{2,T+1} \\ x_{3,T+1} \end{pmatrix}^* \begin{pmatrix} P_{22}^- & P_{23}^- \\ P_{23}^{-*} & P_{33}^+ \end{pmatrix} \begin{pmatrix} x_{2,T+1} \\ x_{3,T+1} \end{pmatrix}. \end{aligned}$$

Moreover, $x_{k+1} = A_{\mathcal{L}}x_k + Bw_k$, so

$$\begin{pmatrix} x_{2,k+1} \\ x_{3,k+1} \end{pmatrix} = \begin{pmatrix} A_{22} & 0 \\ 0 & A'_{33} \end{pmatrix} \begin{pmatrix} x_{2,k} \\ x_{3,k} \end{pmatrix} + \begin{pmatrix} B_2 \\ B_3 \end{pmatrix} w_k.$$

Now $\sigma(A_{22}) \subset \mathcal{T}$, $\sigma(A'_{33}) \subset \mathcal{D}$. By Lemma 4.6 there is $w \in \ell_2$ such that

$$\sum_{k=0}^{\infty} \|w_k\|^2_{R+B^*P_{\mathcal{L}}B} < \epsilon \quad \text{and} \quad \begin{pmatrix} x_2 \\ x_3 \end{pmatrix}(x_0,w)_k \to 0 \quad \text{as } k \to \infty.$$

Introduce

$$u_k = w_k - (R + B^*P_{\mathcal{L}}B)^{-1}(C + B^*P_{\mathcal{L}}A)x_k.$$

Then

$$\begin{aligned} J(x_0,u) &= \lim_{T\to\infty} J_T(x_0,u) \\ &= x_0^* P_{\mathcal{L}} x_0 + \sum_{k=0}^{\infty} \|w_k\|^2_{R+B^*P_{\mathcal{L}}B} \\ &\le \epsilon + x_0^* P_{\mathcal{L}} x_0. \qquad \square \end{aligned}$$

We now turn to the proof of Theorem 2.1.

Proof of Theorem 2.1. Part (i) is proved in Lemmas 4.1, 4.7, and 4.8. It remains to prove (ii).

(ii) First assume that for all x_0 there is an optimal control, i.e., $u^+ \in U_{\mathcal{L}}(x_0)$, for which $V_{\mathcal{L}}(x_0) = J(x_0, u^+)$. Choose x_0 and let u^+ be the corresponding optimal control. Put $x_k^+ = x(x_0,u^+)_k$. By Lemma 4.7

$$x_0^* P_{\mathcal{L}} x_0 = J(x_0,u^+) \ge x_0^* P_{\mathcal{L}} x_0 + \sum_{k=0}^{\infty} \|w_k\|^2_{R+B^*P_{\mathcal{L}}B},$$

where $w_k = u_k^+ + (R+B^*P_{\mathcal{L}}B)^{-1}(C+B^*P_{\mathcal{L}}A)x_k^+$. Hence $w_k = 0$, i.e.,

$$u_k^+ = -(R+B^*P_{\mathcal{L}}B)^{-1}(C+B^*P_{\mathcal{L}}A)x_k^+. \tag{4.7}$$

Since $x_{k+1} = A_{\mathcal{L}} x_k + B w_k$, this yields $x_{k+1}^+ = A_{\mathcal{L}} x_k^+$; in particular,

$$\begin{pmatrix} x_{2,k+1}^+ \\ x_{3,k+1}^+ \end{pmatrix} = \begin{pmatrix} A_{22} & 0 \\ 0 & A_{33}' \end{pmatrix} \begin{pmatrix} x_{2,k}^+ \\ x_{3,k}^+ \end{pmatrix}.$$

As $\sigma(A_{33}') \subset \mathcal{D}$ we have $x_{3,k}^+ \to 0$ as $k \to \infty$. From (4.6) we see that

$$J_T(x_0,u^+) = x_0^* P_{\mathcal{L}} x_0 - x_{3,T+1}^{+*} \Delta_{33} x_{3,T+1}^+ - x_{T+1}^{+*}(P_- - \lambda H^*H)x_{T+1}^+ - \lambda \|Hx_{T+1}^+\|^2.$$

As $J_T(x_0,u^+) - x_0^* P_{\mathcal{L}} x_0 \to 0$, $Hx_{T+1} \to 0$ and $x_{3,T+1}^{+*} \Delta_{33} x_{3,T+1}^+ \to 0$ as $T \to \infty$, we get $x_{T+1}^{+*}(P_- - \lambda H^*H)x_{T+1}^+ \to 0$. As $P_- - \lambda H^*H \le 0$ this gives $(P_- - \lambda H^*H)x_{T+1}^+ \to 0$ and hence $P_- x_T \to 0$. In turn this implies that $D_2 x_{2,k}^+ + D_3 x_{3,k}^+ \to 0$ (see (4.5)). As $x_{3,k} \to 0$ we obtain $D_2 x_{2,k}^+ \to 0$. Using $x_{2,k+1}^+ = A_{22} x_{2,k}^+$ together with the fact that $\sigma(A_{22}) \subset \mathcal{T}$, this yields $D_2 = 0$ (note that x_0, and therefore $x_{2,0}$, is arbitrary). We conclude that

$$\ker \Delta = \mathcal{X}_2 \subseteq \ker \begin{pmatrix} P_- \\ H \end{pmatrix} = \mathcal{L} \cap \ker P_-.$$

Conversely, suppose that $\ker \Delta \subseteq \mathcal{L} \cap \ker P_-$. Then we have $P_{22}^- = 0$ and $P_{23}^- = 0$. Also $D_2 = 0$. Put $u = \{u_k\}$, where u_k is given by

$$u_k = -(R+B^*P_{\mathcal{L}}B)^{-1}(C+B^*P_{\mathcal{L}}A)x_k.$$

Then by (4.6)

$$J_T(x_0,u) = x_0^* P_{\mathcal{L}} x_0 - x_{3,T+1}^* P_{33}^+ x_{3,T+1}.$$

Since $x_{3,k+1} = A'_{33}x_{3,k}$ and $\sigma(A_{33})' \subset \mathcal{D}$, we have $x_{3,T+1} \to 0$. Hence $J_T(x_0, u) \to x_0^* P_{\mathcal{L}} x_0$, so $J(x_0, u) = x_0^* P_{\mathcal{L}} x_0$. Also note that

$$\begin{pmatrix} P_- \\ H \end{pmatrix} x_k = D_3 x_{3,k} \to 0$$

and hence $Hx_{3,k} \to 0$. Thus $u \in U_{\mathcal{L}}(x_0)$ and we can conclude that u is optimal.

The second part of (ii) was already proved (cf. (4.7)) □

REFERENCES

[1] W. A. COPPEL, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377–401.

[2] P. LANCASTER, L. RODMAN, AND A. C. M. RAN, *Hermitian solutions of the discrete algebraic Riccati equation*, Internat. J. Control, 44 (1986), pp. 777–802.

[3] A. C. M. RAN AND L. RODMAN, *Stability of invariant maximal semi-definite subspaces* I, Linear Algebra Appl., 62 (1984), pp. 51–86.

[4] ———, *Stable Hermitian solutions of discrete algebraic Riccati equation*, Math. Control Signals Systems, 5 (1992), pp. 165–193.

[5] H. L. TRENTELMAN, *The regular free-endpoint linear quadratic problem with indefinite cost*, SIAM J. Control Optim., 27 (1989), pp. 27–42.

[6] J. M. SOETHOUDT AND H. L. TRENTELMAN, *The regular indefinite linear-quadratic problem with linear endpoint constraints*, Systems Control Lett., 12 (1989), pp. 23–31.

[7] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[8] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

[9] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Academic Press, San Diego, CA, 1985.

[10] W. W. LIN, *A new method for computing the closed loop eigenvalues of a discrete time algebraic Riccati equation*, Linear Algebra Appl., 96 (1987), pp. 157–180.

[11] V. MEHRMANN, *Existence, uniqueness and stability of solutions to singular linear quadratic optimal control problems*, Linear Algebra Appl., 121 (1989), pp. 291–321.

[12] T. PAPAS, A. J. LAUB, AND N. R. SANDELL, *On the numerical solution of the discrete time algebraic Riccati equation*, IEEE Trans. Automat. Control, Vol. AC-25 (1980), pp. 631–641.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 798–812, July 1993

ERROR BOUNDS FOR THE COMPUTATION OF NULL VECTORS WITH APPLICATIONS TO MARKOV CHAINS*

JESSE L. BARLOW†

Abstract. This paper considers the solution of the homogeneous system of linear equations $Ap = 0$ subject to $c^H p = 1$ where $A \in \mathbf{C}^{n \times n}$ is a singular matrix of rank $n-1$, $c, p \in \mathbf{C}^n$, and $c^H A = 0$. It is assumed that the vector c is known. An important applications context for this problem is that of finding the stationary distribution of a Markov chain. In that context, A is a real singular M-matrix of the form $A = I - Q^T$ where Q is row stochastic. In previous work by Barlow [*SIAM J. Algebraic Discrete Methods*, 7(1986), pp. 414–424], it was shown that, for A an M-matrix, this problem could be solved by solving a nonsingular linear system B of degree $n-1$ that was a principal submatrix of A. Moreover, this matrix B could always be chosen so that $\|B^{-1}\|_2 \approx \|A^\dagger\|_2$. Thus a stable algorithm to solve this problem could be developed for the Markov modeling problem using LU decomposition without pivoting and an update step that identified the correct submatrix.

In this paper, that result is improved in several ways. First, it is shown that many of the results can be generalized to the case where A is any complex matrix of degree n with rank $n-1$; thus they can be generalized to the case where $A = \lambda I - Q^H$ and λ is a simple eigenvalue of Q^H. Second, it is shown that the update step is not necessary for the Markov modeling problem. For that problem, LU decomposition is used without pivoting. Finally, it is shown that a more elegant characterization and sharper error bounds are obtained for this algorithm in terms of the group inverse $A^\#$. In fact, the error is shown to be bounded in terms of $\|A^\#\|_1 \| |A||p| \|_1$, which is smaller than the more traditional condition number $\|A^\#\|_1 \|A\|_1$. These bounds show that the procedure is unconditionally stable.

Key words. group inverse, eigenvectors, condition numbers

AMS subject classifications. 65F05, 65F15, 65F20, 65G05

1. Introduction.

We consider the problem of solving the homogeneous system of linear equations

$$Ap = 0 \tag{1}$$

subject to the constraint

$$c^H p = 1, \tag{2}$$

where $A \in \mathbf{C}^{n \times n}$ is a singular matrix of rank $n-1$, c, $p \in \mathbf{C}^n$, and

$$c^H A = 0. \tag{3}$$

This problem arises naturally if $A = \lambda I - Q^H$ where $\lambda \in C$ is a simple eigenvalue of Q^H. Thus p is a right eigenvector of Q^H and c is a left eigenvector.

An important applications context for (1)–(2) is that of finding the stationary distribution of a Markov chain for which we have the following definition.

DEFINITION 1.1. The *Markov chain problem* is that of finding the vector p in (1)–(2) where $A = I - Q^T$, $c = (1, 1, \ldots, 1)^T$, and Q is row stochastic. For this problem A, c, p, and Q are assumed to be real.

If the chain is ergodic, then Q is irreducible. Also, A will be a singular M-matrix, thus according to the Perron–Frobenius theorem [30, p. 30] or [4, p. 27], p is unique and positive. For practical Markov chain problems, A will be large and sparse [24].

* Received by the editors August 12, 1991; accepted for publication (in revised form) December 30, 1991. This work was supported by the National Science Foundation under grant CCR-9000526.

† Department of Computer Science, The Pennsylvania State University, University Park, Pennsylvania 16802-6103 (barlow@cs.psu.edu).

We consider a simple algorithm to solve (1)–(2) for p. Let A have the form

$$A = P_j \begin{pmatrix} B & y \\ z^H & \alpha_{nn} \end{pmatrix} P_k^T, \tag{4}$$

where P_j exchanges rows j and n, and P_k^T exchanges columns k and n. It is well known that there is always a choice of j and k such that B is nonsingular. Moreover, if A is an irreducible M-matrix or merely irreducible and diagonally dominant, any choice of j and k will make B nonsingular. Also, if $j = k$, then B is a nonsingular M-matrix.

If we know j and k a priori, we can solve for p using the following procedure discussed by Harrod and Plemmons [21], [20] and Barlow [1].

ALGORITHM 1.1.

```
1. Find the permutation matrices $P_j$ and $P_k^T$.
2. Solve $B\hat{x} = -y$ by LU decomposition (or QR decomposition).
3. Let $x = P_k(\hat{x}, 1)^T$.
4. Set $p = x/c^H x$.
```

In [1], it was shown that if A is an M-matrix, then there is always a choice of j and k such that

$$\|A^\dagger\|_2 \leq \|B^{-1}\|_2 \leq (\sqrt{n}+1)^2\|A^\dagger\|_2, \tag{5}$$

where $A^\dagger$ is the Moore–Penrose inverse of A [16, p. 243]. Thus, from the well-known fact that $\|B\|_2 \leq \|A\|_2$, we have that

$$\kappa_2(B) \leq (\sqrt{n}+1)^2 \kappa_2(A), \tag{6}$$

$$\kappa_2(B) = \|B\|_2\|B^{-1}\|_2, \qquad \kappa_2(A) = \|A\|_2\|A^\dagger\|_2.$$

This result explained experiments in Harrod and Plemmons [21]. Chu [8] points out that the factor n can be large, but, in §3, we show that, contrary to statements in that paper, this does not adversely affect the stability of Algorithm 1.1.

In §2, we generalize this result to an arbitrary problem of the form (1)–(2) and to any generalized inverse A^- satisfying

$$A = AA^-A. \tag{7}$$

A particularly important matrix from the class (7) is the group inverse, denoted $A^\#$. It is the unique matrix such that

$$1.\ AA^\#A = A, \quad 2.\ A^\#AA^\# = A^\#, \quad 3.\ AA^\# = A^\#A. \tag{8}$$

The group inverse exists if and only if $\text{rank}(A) = \text{rank}(A^2)$ and the latter condition holds since zero is a simple eigenvalue of A. As shown by Meyer [25], it yields a more elegant characterization of the problem (1)–(2) than does the Moore–Penrose inverse. From Meyer [25], we note that $A^\dagger = A^\#$ if and only if $c = p$.

In [1], we showed that for the Markov modeling problem, it is easy to choose j and k (with $j = k$), and Algorithm 1.1 would just have to be followed by an update step such as in [13] after choosing the value of k. We show here that the update step can be avoided. The results in §2 are proven using different side conditions for (1). General side conditions for (1) are discussed by Meyer and Stewart [26].

In §3, we show that, with a slight modification, Algorithm 1.1 is stable for the Markov chain problem and for any problem of the form (1)–(2) as long as

1. The left eigenvector c is known;
2. The right eigenvector p is simple.

These results are characterized in terms of the group inverse and two measures of the condition in the 1-norm. The first is the "group" condition number

$$\kappa_1^{\#}(A) = \|A\|_1 \|A^{\#}\|_1, \tag{9}$$

and the second is the "effective" condition number

$$\kappa_1^{E}(A) = \|A^{\#}\|_1 \||A||p|\|_1. \tag{10}$$

Throughout the paper $|A|$ denotes the matrix $|A| = (|a_{ij}|)$ and $|p|$ is the vector $|p| = (|p_1|, |p_2|, \ldots, |p_n|)^T$. Since $\|p\|_1 = 1$ for the Markov modeling problem, we have $\kappa_1^E(A) \leq \kappa_1^{\#}(A)$. From our tests in §4, $\kappa_1^E(A)$ seems to be a more reliable estimate of the condition of the problem (1)–(2). Good estimates of $\kappa_1^{\#}(A)$ and $\kappa_1^E(A)$ can be obtained using the Hager–Higham [19], [22] method. A condition number of this form for the problem (1)–(2) was introduced by Geurts [15].

The term "effective" condition number is similar to a term used by Chan and Fousler [7] and has similar implications here, but our definition is slightly different from theirs.

As shown by Harrod and Plemmons [21], we can use Gaussian elimination without pivoting [12], which was first advocated for this problem by Funderlic and Mankin [10]. Golub and Meyer [17] discuss the use of orthogonal factorization for dense matrices. It is easy to use standard sparse matrix software such as SPARSPAK-A or the Harwell code MA28 when A is large and sparse. Static storage allocation can be used throughout the computation.

Contrary to statements by Chu [8], who advocates a more expensive bordering strategy along with a rank revealing orthogonal decomposition [3], [9], [6], [5] for this problem, the fact that we can use Gaussian elimination without pivoting represents significant savings in storage. For large sparse systems, orthogonal decomposition involves considerably more storage and operations [14].

Section 5 gives the conclusions to our paper.

2. Perturbation results for the problem.

Suppose that we have obtained an approximate solution $\bar{p}$ to (1)–(2) by some numerical method. Using the techniques of backward error analysis, we have that $\bar{p}$ satisfies

$$A\bar{p} = r, \tag{11}$$

$$c^H \bar{p} = 1 + \epsilon, \tag{12}$$

where r and ϵ are the residual errors in computation. Although A is singular, (13) is clearly consistent. Thus, if we let $\Delta p = \bar{p} - p$, then

$$A\Delta p = r, \tag{13}$$

$$c^H \Delta p = \epsilon \tag{14}$$

is consistent.

All solutions of (13) can be written

$$\Delta p = A^{-}r + \alpha p, \tag{15}$$

where α is a scalar. This scalar is determined by side conditions of the form (14). However, if $A^{-} = A^{\#}$, and if we use the condition (14), then

$$c^H \Delta p = c^H A^{\#} r + \alpha c^H p = \epsilon.$$

Since $r = A\bar{p}$ and $c^H p = 1$, we have

$$\begin{aligned}\epsilon &= c^H A^{\#} A \bar{p} + \alpha c^H p \\ &= c^H A A^{\#} \bar{p} + \alpha = \alpha\end{aligned}$$

since $c^H A = 0$. Therefore,

$$\Delta p = A^{\#} r + \epsilon p. \tag{16}$$

Thus the solution to (13)–(14) is quite elegant in terms of the group inverse. We now give a lemma that uses (15) and (16) to relate the Moore–Penrose inverse and the group inverse. We present them without proof; their proofs are simple and contained in [2].

LEMMA 2.1. *For every $r \in \text{Range}(A)$*

$$A^{\#} r = A^{\dagger} r - (c^H A^{\dagger} r) p, \tag{17}$$

$$A^{\dagger} r = A^{\#} r - \frac{(p^H A^{\#} r)}{p^H p} p. \tag{18}$$

For the Markov chain problem, we have that

$$c^H p = \|c\|_\infty \|p\|_1 = 1.$$

Since $c = e = (1, 1, \ldots, 1)^T$ and $p \geq 0$. Hence we have the following corollaries bounding the norm of $A^{\dagger} r$ and $A^{\#} r$.

COROLLARY 2.2. *If $c^H p = \|c\|_\infty = \|p\|_1 = 1$ then for each $r \in \text{Range}(A)$, we have*

$$\frac{1}{1+\sqrt{n}} \|A^{\dagger} r\|_1 \leq \|A^{\#} r\|_1 \leq 2 \|A^{\dagger} r\|_1. \tag{19}$$

In the next section, we use a similar technique to analyze Algorithm 1.1 and to generalize results in [1].

3. Analysis of the algorithm to compute the null vector.

In this section, we will show that Algorithm 1.1 obtains a solution to (1)–(2) that is as good as can be expected. We begin our analysis of the algorithm with a technical lemma that will shorten many of our arguments.

LEMMA 3.1. *Let $\tilde{x} \in \mathbf{C}^{n-1}$ be a vector that satisfies*

$$B\tilde{x} = -y + r_B. \tag{20}$$

Let A have the form (4). Let $c = (c_1, \ldots, c_{n-1}, c_n)^T$ satisfy (3), assume $c_n \neq 0$, and define $d = -c_n^{-1}(c_1, \ldots, c_{n-1})^T$. Let $\|\cdot\|_s$ and $\|\cdot\|_q$ be Hölder norms. Then $\bar{x} = (\tilde{x}^T, 1)^T$ satisfies

$$A\bar{x} = r_A, \tag{21}$$

where

$$\|r_A\|_s \leq (1+\|d\|_q)\|r_B\|_s, \qquad s^{-1}+q^{-1}=1. \tag{22}$$

Proof. The hypothesis implies that

$$(B\;y)\bar{x} = r_B.$$

Thus

$$A\bar{x} = \begin{pmatrix} r_B \\ \rho_B \end{pmatrix}.$$

Since $c^H A = 0$, we have

$$c^H \begin{pmatrix} r_B \\ \rho_B \end{pmatrix} = 0$$

so that

$$\rho_B = d^H r_B. \tag{23}$$

Thus by Hölder's inequality

$$\begin{aligned} \|A\bar{x}\|_s &\leq \|r_B\|_s + |\rho_B| \\ &\leq \|r_B\|_s + |d^H r_B| \\ &\leq (1+\|d\|_q)\|r_B\|_s. \qquad \square \end{aligned}$$

We now give a lemma on the perturbation theory of the system (1).

LEMMA 3.2. *Assume the hypothesis and notation of Lemma* 3.1 *and let* $\Delta x = (\tilde{x}-\hat{x},0)^T$ *where* $\hat{x}$ *solves* $B\hat{x}=-y$. *Assume also* $p_n \neq 0$ *where* $p=(p_1,\ldots,p_n)^T$ *is the solution to* (1)–(2). *Then*

$$\|\Delta x\|_s = \|B^{-1}r_B\|_s \leq \left(1+\frac{\|p\|_s}{|p_n|}\right)(1+\|d\|_q)\|A^-\|_s\|r_B\|_s\; s^{-1}+q^{-1}=1, \tag{24}$$

where A^- *is any matrix satisfying* (7).

Proof. We have that $\bar{x}=(\tilde{x},1)^T$ solves

$$A\bar{x}=r_A, \qquad e_n^H\bar{x}=1,$$

where $e_n=(0,\ldots,1)^T$, and from Lemma 3.1, r_A satisfies (22). Clearly, $x=(\hat{x},1)^T$ solves

$$Ax=0, \qquad e_n^H x=1,$$

so that $\Delta x = \bar{x}-x = (\tilde{x}-x,0)^T = (B^{-1}r_B,0)^T$ solves

$$A\Delta x = r_A, \qquad e_n^H\Delta x=0.$$

Thus

$$\Delta x = A^- r_A - \frac{(e_n^H A^- r_A)}{e_n^H p}p,$$

where A^- is any matrix satisfying (7). Thus

$$\|\Delta x\|_s \leq \|A^- r_A\|_s + \frac{\|e_n\|_q \|A^- r_A\|_s \|p\|_s}{|p_n|} \tag{25}$$

$$\leq \left(1 + \frac{\|p\|_s}{|p_n|}\right) \|A^- r_A\|_s \leq \left(1 + \frac{\|p\|_s}{|p_n|}\right) \|A^-\|_s \|r_A\|_s. \tag{26}$$

The use of norm inequalities and equation (22) establishes the lemma. □

There is an immediate corollary to this result that is relevant to the Markov chain problem.

COROLLARY 3.3. *Let P_j be the permutation matrix that exchanges rows j and n of A. Then there exist j and k such that if*

$$A = P_j \begin{pmatrix} B & y \\ z^H & \alpha_{nn} \end{pmatrix} P_k^T,$$

then

$$\|B^{-1}\|_1 \leq 2(1+n)\|A^-\|_1, \tag{27}$$

$$\|B^{-1}\|_2 \leq (1+\sqrt{n})(1+\sqrt{n-1})\|A^-\|_2, \tag{28}$$

for any matrix A^- satisfying (7).

Proof. Simply choose j such that $|c_j| = \|c\|_\infty$ and k such that $|p_k| = \|p\|_\infty$ and apply Lemma 3.2. □

If we take $A^- = A^\dagger$, then (28) is a generalization of the bound (5) from [1].

For the Markov chain problem, $\|p\|_1 = 1$ and $c = e = (1, 1, \ldots, 1)^T$, so these conditions are always met. In general, this gives us

$$\|B^{-1}\|_1 \leq 2\left(1 + \frac{1}{|p_n|}\right) \|A^-\|_1. \tag{29}$$

Algorithm 1.1 actually solves (1)–(2) more accurately than Lemma 3.2 would suggest. The following analysis indicates this. First, we begin with a technical lemma.

LEMMA 3.4. *Let $\bar{p} \in \mathbf{C}^n$ be the floating point computation of*

$$\bar{p} = fl\left(\frac{1}{c^H \bar{x}} \bar{x}\right), \tag{30}$$

where $c, \bar{x} \in \mathbf{C}^n$. Then

$$\bar{p} = \frac{1+\epsilon_1}{c^H \bar{x}} D\bar{x}, \tag{31}$$

where

$$|\epsilon_1| \leq (n + 2\sqrt{2} - 1)\|c\|_\infty \|p\|_1 u + O(u^2), \tag{32}$$

$$\|D - I\|_1 \leq 5\sqrt{2}u + O(u^2). \tag{33}$$

Proof. From [33, p. 447], for two complex numbers ζ_1 and ζ_2, we have

$$fl(\zeta_1 \pm \zeta_2) = (\zeta_1 \pm \zeta_2)(1+\delta_1), \qquad |\delta_1| \le u, \tag{34}$$

$$fl(\zeta_1\zeta_2) = \zeta_1\zeta_2(1+\delta_2), \qquad |\delta_2| \le 2\sqrt{2}u + O(u^2), \tag{35}$$

$$fl(\zeta_1/\zeta_2) = (\zeta_1/\zeta_2)(1+\delta_3), \qquad |\delta_3| \le 5\sqrt{2}u + O(u^2). \tag{36}$$

Here δ_i, $i = 1, 2, 3$, are complex numbers.

Clearly, (34) and (35), plus standard results for inner products, yield the bound

$$|fl(c^H\bar{x}) - c^H\bar{x}| \le (n + 2\sqrt{2} - 1)\|c\|_\infty \|\bar{x}\|_1 u + O(u^2). \tag{37}$$

If we divide by $c^H\bar{x}$, we obtain

$$\frac{|fl(c^H\bar{x}) - c^H\bar{x}|}{|c^H\bar{x}|} \le (n + 2\sqrt{2} - 1)\|c\|_\infty \|p\|_1 u + O(u^2). \tag{38}$$

Thus

$$\frac{1}{fl(c^H\bar{x})} = \frac{1+\epsilon_1}{|c^H\bar{x}|}, \qquad |\epsilon_1| \le (n + 2\sqrt{2} - 1)\|c\|_\infty \|p\|_1 u + O(u^2).$$

The complex division yields the result

$$fl\left(\frac{\bar{x}}{c^H\bar{x}}\right) = \frac{1+\epsilon_1}{c^H\bar{x}} D\bar{x},$$

where $\|D - I\|_1 \le 5\sqrt{2}u + O(u^2)$. □

We can now prove the theorem that establishes the stability of Algorithm 1.1.

THEOREM 3.5. *Let $\tilde{x}$ be the solution to* (20) *and let $\bar{p}$ be the computed result of Algorithm* 1.1 *from* (30) *with $\bar{x} = (\tilde{x}^T, 1)^T$. Assume the hypothesis and notation of Lemma* 3.1 *and that r_B in* (20) *satisfies*

$$\|r_B\|_1 \le \psi(n)\|A\|_1\|\bar{x}\|_1 u + O(u^2), \tag{39}$$

where $\psi(n)$ is a modestly growing function of n. Let p be the exact solution of (1)–(2), *then*

$$\frac{\|\Delta p\|_1}{\|\bar{p}\|_1} \le [(1 + \|d\|_\infty)\psi(n) + 5\sqrt{2}]\kappa_1^{\#}(A) \tag{40}$$

$$+(n + 2\sqrt{2})\|c\|_\infty \|p\|_1 u + O(u^2), \tag{41}$$

where $\kappa_1^{\#}(A) = \|A\|_1\|A^{\#}\|_1$ is the "group" condition number. If $\|r_B\|_1$ satisfies

$$\|r_B\|_1 \le \phi(n)\||A||\bar{x}|\|_1 u + O(u^2), \tag{42}$$

then

$$\|\Delta p\|_1 \le ([1 + \|d\|_\infty]\phi(n) + 5\sqrt{2})\kappa_1^{E}(A)u \tag{43}$$

$$+(n + 2\sqrt{2})\|c\|_\infty \|p\|_1 u + O(u^2), \tag{44}$$

where $\kappa_1^E(A)$ *is as defined in* (10).

Proof. From the hypothesis, we have that

$$\|(B\,y)\bar{x}\|_1 = \|r_B\|_1 \leq \psi(n)\|A\|_1\|\bar{x}\|_1 u + O(u^2).$$

From Lemma 3.4, we have

$$c^H\bar{p} = 1 + \epsilon_2 \tag{45}$$

and

$$\|r_A\|_1 \leq \|A\bar{p}\|_1 = \frac{1+\epsilon_1}{|c^H\bar{x}|}\|AD\bar{x}\|_1,$$

where

$$|\epsilon_1| \leq n\|c\|_\infty\|\bar{p}\|_1 u + O(u^2)$$

and D is a diagonal matrix such that

$$\|D - I\|_1 \leq 5\sqrt{2}u + O(u^2).$$

Thus

$$\|r_A\|_1 = \frac{1+\epsilon_1}{|c^H\bar{x}|}[\|A\bar{x}\|_1 + \|A(D-I)\bar{x}\|_1].$$

The first term is bounded using Lemma 3.1, and by standard norm inequalities, we have

$$\|r_A\|_1 \leq [(1+\|d\|_\infty)\psi(n) + \|D-I\|_1]\|A\|_1\left[\frac{1+\epsilon_1}{|c^H\bar{x}|}\|\bar{x}\|_1\right] + O(u^2). \tag{46}$$

We note that, because of the bound on D and the fact that u is the machine unit, then D is nonsingular and $\|D^{-1}\|_1 \leq 1 + 5\sqrt{2}u + O(u^2)$. Thus

$$\frac{1+\epsilon_1}{|c^H\bar{x}|}\|\bar{x}\|_1 = \|D^{-1}\bar{p}\|_1 \leq \|D^{-1}\|_1\|p\|_1 \leq \|p\|_1 + O(u). \tag{47}$$

Combining (46) and (47) yields

$$\|r_A\|_1 \leq [(1+\|d\|_\infty)\psi(n) + 5\sqrt{2}]\|A\|_1\|\bar{p}\|_1 u + O(u^2).$$

Now to bound the error in p. Using (15), and taking norms, we have that

$$\|\bar{p} - p\|_1 \leq \|A^{\#}\|_1\|r_A\|_1 + |\epsilon_2|\|p\|_1.$$

Using the bound for $\|r_A\|_1$ and $|\epsilon_2|$ and dividing by $\|\bar{p}\|_1$ yields (40).

A similar derivation with the assumption (42) for $\|r_B\|_1$ allows one to substitute to obtain (43). □

Remark 1. Meyer and Stewart [26] describe a strong relationship between $\|A^{\#}\|_2$ and the sep(·) function that is commonly used to bound the error in eigenvectors [28]. All of the bounds given in this paper can be easily reworked in terms of sep(·), but they are somewhat less elegant.

Remark 2. Funderlic and Meyer [11] use the condition number $\max_{(i,j)} |a^{\#}_{ij}|$. The result here can be made consistent with theirs from the observation that

$$\begin{aligned}\|\Delta p\|_\infty &\leq \|A^{\#}r\|_\infty + \epsilon\|p\|_\infty \\ &\leq \|A^{\#}\|_{\infty,1}\|r\|_1 + |\epsilon|\|p\|_\infty,\end{aligned}$$

where $\|A^{\#}\|_{\infty,1} = \max_{\|x\|_1=1} \|A^{\#}x\|_\infty = \max_{1\leq i,j\leq n} |a^{\#}_{ij}|$. Thus we can obtain bounds on $\|\Delta p\|_\infty$ instead of $\|\Delta p\|_1$ and define condition numbers $\kappa^{\#}_{\infty,1}(A) = \|A^{\#}\|_{\infty,1}\|A\|_1$ and $\kappa^{E}_{\infty,1}(A) = \|A\|_{\infty,1}\||A||p|\|_1$. The comparisons between these two condition numbers will be identical to those between $\kappa^{\#}_1(A)$ and $\kappa^{E}_1(A)$. We choose the latter because known software [19], [23] can be used to estimate them [2]. Also, this keeps as many of the bounds as possible in the same family of norms.

A similar line of analysis can obtain the componentwise bound

$$\begin{aligned}|(\Delta p)_i| \leq ([1 + \|d\|_\infty]\phi(n) + 5\sqrt{2}) \max_{1\leq j\leq n} |a^{\#}_{ij}|\||A||p|\|_1 \\ + (n + 2\sqrt{2})\|c\|_\infty|p_i| + O(u^2), \qquad i = 1, 2, \ldots, n\end{aligned}$$

under the assumption (42).

Since $c = (1, 1, \ldots, 1)^T$ for the Markov chain problem, the bound in Theorem 3.5 simplifies. The corollary below summarizes a simple version of this bound.

COROLLARY 3.6. *Assume the hypothesis of Theorem* 3.5. *If A arises out of the Markov chain problem in Definition* 1.1, *then*

$$\|\Delta p\|_1 \leq [2\psi(n) + 5\sqrt{2}]\kappa^{\#}_1(A)u + (n + 2\sqrt{2})u + O(u^2).$$

Under the assumption (42), *we have that*

$$\|\Delta p\|_1 \leq [2\phi(n) + 5\sqrt{2}]\kappa^{E}_1(A)u + (n + 2\sqrt{2})u + O(u^2).$$

The above results mean that for any problem of the form (1)–(2) where the right null vector p is simple and the left null vector c is known, we need only be concerned about the accuracy of solving the nonsingular linear system

$$B\hat{x} = -y. \tag{48}$$

That problem is considered below.

The recommendations in the remainder of this section concern only the Markov modeling problem where A is real and has the form $A = I - Q^T$, where Q is row stochastic and is thus diagonally dominant. In that case

$$B = (b_{ij}),$$

where $b_{jj} > 0$ and $b_{ij} \leq 0, i \neq j$; thus B is a nonsingular M-matrix. Varga and Cai [31] show that we can perform Gaussian elimination without pivoting and the diagonal entries b_{jj} should remain positive while the off-diagonal entries $b_{ij}, i \neq j$ remain nonpositive throughout the computation. Thus, if for some k, b_{kk} becomes negative or zero during the course of the algorithm, then B must be singular to machine precision.

Since B is diagonally dominant, combining results of Wilkinson [32] and Reid [27], we have that Gaussian elimination without pivoting yields an LU decomposition such that

$$L(U\,\tilde{y}) = (B\,y) + E, \qquad E = (e_{ij}),$$

where

$$|e_{ij}| \leq 6.02|a_{jj}|m_{ij}u, \qquad i = 1, 2, \ldots, n-1, \qquad j = 1, 2, \ldots, n.$$

Here m_{ij} is the number of operations performed on the entry b_{ij}. Thus Gaussian elimination on B satisfies a bound of the form (39). Neglecting the back substitution errors which satisfy similar bounds [33], we have

$$(B\,y)p = r_B = -Ep.$$

We can now say that

$$\|r_B\|_1 = \|Ep\|_1 \leq 6.02\phi_0(n)\|\mathrm{diag}(a_{11}, \ldots, a_{nn})p\|_1 u, \tag{49}$$

where

$$\phi_0(n) = \max_{(i,j)} m_{ij}n_j, \tag{50}$$

$$n_j = \text{ number of nonzeros in column } j \text{ of } L \text{ and } U. \tag{51}$$

We can now give a theorem that shows that Algorithm 1.1 obtains answers that are as good as can be expected for the Markov modeling problem.

THEOREM 3.7. *Let $A = I - Q^T$ where Q is an irreducible row stochastic matrix. Let $Ap = 0$ be solved by Algorithm 1.1 where the system $Bx = -y$ is solved by Gaussian elimination without pivoting. Neglecting the error from back substitution, if no diagonal entry of B becomes negative or zero, then*

$$\|\Delta p\|_1 \leq \ [\phi_0(n) + 5\sqrt{2}]\kappa_1^E(A)u \tag{52}$$

$$+ (n + 2\sqrt{2})u + O(u^2), \tag{53}$$

where

$$\phi_0(n) = 6.02 \max_{(i,j)} m_{ij}n_j,$$
$$m_{ij} = \textit{number of operations performed on entry } a_{ij},$$
$$n_j = \textit{number of nonzeroes in column } j \textit{ of } L \textit{ and } U.$$

Proof. It suffices to show that

$$\|r_B\|_1 \leq \tfrac{1}{2}\phi_0(n)\||A||p|\|_1 u. \tag{54}$$

We then use Theorem 3.5.

The above analysis yields (49). Let $\tilde{D} = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn})$. We have that

$$|A||p| = |\tilde{D}p| + |A - \tilde{D}||p|. \tag{55}$$

From $Ap = 0$ and the fact that $p \geq 0$, $\tilde{D} \geq 0$, and $A - \tilde{D} \leq 0$, we have

$$\tilde{D}p = -(A - \tilde{D})p \geq 0. \tag{56}$$

TABLE 1
Residuals and forward error for random matrices.

Quantity	Maximum	Average
$\|\|r_B\|\|_1$	$1.663E{-13}$	$1.307E{-13}$
$\|\|r_A\|\|_1$	$1.884E{-13}$	$1.450E{-13}$
$\|\|\Delta p\|\|_1$	$8.313E{-7}$	$2.859E{-7}$

Combining (55) and (56) yields

$$|A||p| = 2|\tilde{D}p|.$$

Thus from (49), we get (52). □

We consider the implementation of Algorithm 1.1 for the Markov modeling problem. In the general case, we recommend the use of the orthogonal factorization approach of Golub and Meyer [17].

The main remaining difficulty in solving (48) is if B turns out to be singular to machine precision. For the solution of (48), we have that B is a nonsingular M-matrix. Thus $B = (b_{ij})$ where $b_{jj} > 0$ and $b_{ij}, i \neq j$, and moreover, in exact arithmetic, these entries will not change sign in the course of elimination [31]. However, if B is ill conditioned, a diagonal entry could become negative or zero because of rounding errors. If that occurs (if, say, $b_{kk} = a_{kk} \leq 0$), then we can use the fact that $a_{kk} = \sum_{i=k+1}^{n} a_{ik}$ when it is time to eliminate column k. This technique was introduced by Grassman, Taksar, and Heyman [18]. A special case of that method for nearly uncoupled problems was analyzed by Stewart and Zhang [29]. We then proceed with the elimination. Thus breakdown in Gaussian elimination can be completely avoided and the error bounds given here are unconditional.

4. Tests using the algorithm. The test program was written in FORTRAN 77 and the test runs were done on a SUN 4 in the author's office at the Pennsylvania State University. The estimated machine precision was 5.960×10^{-8}. We did one large random test set plus two particular ill-conditioned problems.

Example 4.1. We tested Algorithm 4.1 with 500 randomly generated dense 50×50 matrices. These matrices were of the form

$$a_{jj} = 0.8e^{(j-1)\omega}, \quad \omega = ln(10^{-7})/49, \quad j = 1, 2, \ldots, n,$$

$$a_{ij} = -\alpha_{ij} \cdot a_{jj}/\beta_j,$$

where $\beta_j = \sum_{i \neq j} \alpha_{ij}$ and α_{ij} are random numbers between 0 and 1. This method gives a random matrix A such that A is a singular irreducible M-matrix. This always insured that $||A||_1 = 1.6$.

Using the Hager–Higham [19], [22] method as discussed in [2], our estimates of $||A^{\#}||_1$ ranged from 3.309×10^6 to 1.136×10^7. In the traditional sense, these matrices are very ill conditioned. According to the $\kappa_1^{\#}(A)$ bound, one should expect $||\Delta p||_1$ to be about 10^{-1} or 10^{-2} However, the $\kappa_1^{E}(A)$ estimates are much smaller. They ranged from 7.431 to 25.30; thus Algorithm 4.1 should obtain accurate answers, which it did. This is demonstrated in Tables 1 and 2.

Here

$$\text{est.A} = ||r_A||_1 * \text{cond.est.}, \tag{57}$$

TABLE 2
Ratios of residuals and estimates to actual errors.

Quantity	Maximum	Average	Minimum
$\frac{\Vert r_A\Vert_1}{\Vert r_B\Vert_1}$	1.392	1.112	1.001
$\frac{\Vert \Delta p\Vert_1}{\text{est.A}}$	1.415	$5.343E{-}1$	$5.524E{-}2$
$\frac{\Vert \Delta p\Vert_1}{\text{est.B}}$	1.970	$6.239E{-}1$	$5.585E{-}2$
$\frac{\Vert \Delta p\Vert_1}{\text{est.Z}}$	1.880	$6.094E{-}1$	$6.354E{-}2$

$$\text{est.B} = \Vert r_B \Vert_1 * \text{cond.est.}, \tag{58}$$

$$\text{est.Z} = \Vert\, |A||p|\, \Vert_1 * \text{cond.est.} * u, \tag{59}$$

where *cond.est.* is the Hager [19] estimate of $\Vert A^{\#}\Vert_1$ and u is the estimated machine precision. Thus we note that both est. A and est. B are good estimates of $\Vert \Delta p\Vert_1$. None of the 500 linear systems were determined to have a rank deficiency of greater than one.

The remaining two examples are taken from Harrod and Plemmons [21]. For both examples, est. A, est. B, and est. Z are as defined in (57), (58), and (59).

Example 4.2.

$$A = \begin{pmatrix} 1.0E-06 & -0.4 & -5.0E-07 & -5.0E-07 & -2.0E-07 \\ -1.0E-07 & 0.7 & 0 & 0 & -3.0E-07 \\ -2.0E-07 & 0 & 1.0E-06 & 0 & -1.0E-07 \\ -3.0E-07 & 0 & 0 & 1.0E-06 & -4.0E-07 \\ -4.0E-07 & -0.3 & -5.0E-07 & -5.0E-07 & 1.0E-06 \end{pmatrix}.$$

We obtained the following computational results.

$$\Vert A\Vert_1 = 1.4, \qquad \Vert A^{\dagger}\Vert_2^{-1} = 1.007E-06.$$

The value for $\Vert A^{\dagger}\Vert_2^{-1}$ is from [21].

The Hager estimate of $\Vert A^{\#}\Vert_1 = 4.444E + 05$. Thus, from the $\kappa_1^{\#}(A)$ estimate one should expect $\Vert \Delta p\Vert_1$ to be about 10^{-2} or 10^{-3}. However, the Hager estimate of $\kappa_1^{E}(A)$ is only 1.011; thus we should expect, and do obtain, good answers.

$$\begin{aligned} &\Vert \Delta p\Vert_1 = 2.061E-08; \\ &\Vert r_B\Vert_1 = 4.441E-14, \qquad \Vert r_A\Vert_1 = 6.561E-14; \\ &\Vert r_A\Vert_1/\Vert r_B\Vert_1 = 1.489; \\ &\Vert \Delta p\Vert_1/\text{est.A} = 0.7068, \qquad \Vert \Delta p\Vert_1/\text{est.B} = 1.052; \\ &\Vert \Delta p\Vert_1/\text{est.Z} = 0.4952. \end{aligned}$$

The solution p is

$$p = (0.315217, 1.95652E-07, 9.81884E-02, 0.235145, 0.351449)^T.$$

Example 4.3. Let

$$\bar{Q}(\epsilon) = \begin{pmatrix} Q_{11} & 0 \\ 0 & Q_{22} \end{pmatrix} + \epsilon(e_1 e_6^T + e_6 e_1^T),$$

where

$$Q_{11} = \begin{pmatrix} 0.1 & 0.3 & 0.1 & 0.2 & 0.3 \\ 0.2 & 0.1 & 0.1 & 0.2 & 0.4 \\ 0.1 & 0.2 & 0.2 & 0.4 & 0.1 \\ 0.4 & 0.2 & 0.1 & 0.2 & 0.1 \\ 0.6 & 0.3 & 0.0 & 0.0 & 0.1 \end{pmatrix}$$

$$Q_{22} = \begin{pmatrix} 0.1 & 0.2 & 0.2 & 0.4 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.3 & 0.2 \\ 0.1 & 0.5 & 0.0 & 0.2 & 0.2 \\ 0.5 & 0.2 & 0.1 & 0.0 & 0.2 \\ 0.1 & 0.2 & 0.2 & 0.3 & 0.2 \end{pmatrix}.$$

Then $A = I - Q^T$ where $Q = D\bar{Q}(\epsilon)$ and $D = \text{diag}(d_1, d_2, \ldots, d_{10})$ is chosen so that Q satisfies $Qc = c$, $c = (1, 1, \ldots, 1)^T$. This is a very ill conditioned problem of a Markov chain with two loosely coupled sets of states. We chose $\epsilon = 1.0E - 06$.

The Hager estimate of $\|A^{\#}\|_1 = 2.278E + 06$. Thus, according to the $\kappa_1^{\#}(A)$ estimate, one should expect $\|\Delta p\|_1$ to be about 10^{-1} or 10^{-2}. In this case, that is accurate, but that is because the $\kappa_1^E(A)$ is estimated to be $4.032E + 06$, so it is not much more optimistic.

$$\|\Delta p\|_1 = 8.423E - 01;$$
$$\|r_B\|_1 = 4.504E - 08, \qquad \|r_A\|_1 = 6.891E - 08;$$
$$\|r_A\|_1/\|r_B\|_1 = 1.530;$$
$$\|\Delta p\|_1/est.A = 0.5365;$$
$$\|\Delta p\|_1/est.B = 0.8210;$$
$$\|\Delta p\|_1/est.Z = 0.3505.$$

The solution p is

$$p = (0.124286, 9.87915E - 02, 3.71796E - 02, 7.43592E - 02, 9.772923E - 02, \\ 0.124286, 0.135159, 7.20948E - 02, 0.135012, 0.101102)^T.$$

The above tests confirm that this algorithm will always yield a small residual for the problem (1)–(2). Note that although $\kappa_1^{\#}(A)$ is large for all of the above problems, only the last example obtains poor results. That is because it is the only one for which $\kappa_1^E(A)$ is large.

We note that the last example is a loosely coupled Markov chain. Stewart and Zhang [29] describe a variant of Gaussian elimination that obtains better error bounds for this particular class of problems. That procedure has a more restrictive pivoting strategy.

5. Conclusion. The partition algorithm described by Harrod and Plemmons [21] is stable provided that we use the diagonal adjustment technique of Grassman, Taksar, and Heyman [18]. Thus for the large sparse problem that normally arises in queueing models, we can fully exploit the nonzero structure.

Error bounds in terms of $\|A^{\#}\|_1 \| |A||p| \|_1$ seem to give very accurate estimates of the error in computing the null vector p.

Acknowledgments. The author would like to thank Bob Plemmons for his careful reading of the manuscript and helpful comments. He also thanks the referee who pointed out reference [11] and made other helpful comments that resulted in making the manuscript more readable. Lastly, he acknowledges the helpful comments of Guodang Zhang, Bob Funderlic, Daniel Heyman, and Francoise Chatelin.

REFERENCES

[1] J. L. Barlow, *On the smallest positive singular value of a singular M-matrix with applications to ergodic Markov chains*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 414–424.

[2] ———, *Error bounds and condition estimates for the computation of null vectors with applications to Markov chains*, Tech. Rep. CS-91-20, Dept. of Computer Science, The Pennsylvania State Univ., University Park, PA, 1991.

[3] J. L. Barlow and U. B. Vemulapati, *Rank detection methods for sparse matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1279–1297.

[4] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[5] C. H. Bischof, J. G. Lewis, and D. J. Pierce, *Incremental condition estimation for sparse matrices*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 644–662.

[6] T. F. Chan, *Rank revealing* QR *factorization*, Linear Algebra Appl., 88/89 (1987), pp. 67–82.

[7] T. F. Chan and D. E. Fousler, *Effectively well-conditioned linear systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 963–969.

[8] K. W. E. Chu, *Bordered matrices, singular systems, and ergodic Markov chains*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 688–701.

[9] L. V. Foster, *Rank and null space calculations using matrix decompositions without column pivoting*, Linear Algebra Appl., 74 (1986), pp. 47–72.

[10] R. E. Funderlic and J. B. Mankin, *Solution of homogeneous systems of linear equations arising in compartmental models*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 375–383.

[11] R. E. Funderlic and C. D. Meyer, Jr., *Sensitivity of the stationary distribution for an ergodic Markov chain*, Linear Algebra Appl., 76 (1986), pp. 1–17.

[12] R. E. Funderlic and R. J Plemmons, LU *decomposition of M-matrices by elimination without pivoting*, Linear Algebra Appl., 41 (1981), pp. 99–110.

[13] ———, *Updating* LU *factorizations for computing stationary distributions*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 30–42.

[14] J. A. George, J. W. H. Liu, and E. Ng, *A data structure for sparse* QR *and* LU *factors*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 877–898.

[15] A. J. Geurts, *A contribution to the theory of condition*, Numer. Math., 39 (1982), pp. 85–96.

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

[17] G. H. Golub and C. D. Meyer, Jr., *Using the* QR *factorization and group inverse to compute, differentiate and estimate the sensitivity of stationary probabilities of Markov chains*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 273–281.

[18] W. K. Grassman, M. I. Taksar, and D. P. Heyman, *Regenerative analysis and steady state distribution for Markov chains*, Oper. Res., 33 (1985), pp. 1107–1116.

[19] W. W. Hager, *Condition estimates*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 311–316.

[20] W. J. Harrod, *Rank modification methods for certain systems of linear equations*, Ph.D. thesis, University of Tennessee, Knoxville, TN, Department of Mathematics, 1984.

[21] W. J. Harrod and R. J. Plemmons, *Comparison of some direct methods for computing stationary distributions of Markov chains*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 453–469.

[22] N. J. Higham, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–598.

[23] ———, *Fortran codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation*, ACM Trans. Math. Software, 14 (1988), pp. 381–396.

[24] L. C. Kaufman, *Matrix methods for queueing problems*, SIAM J. Sci. Statist. Computing, 4 (1983), pp. 525–552.

[25] C. D. Meyer, Jr., *The role of the group inverse in the theory of finite Markov chains*, SIAM Rev., 17 (1975), pp. 443–464.

[26] C. D. Meyer, Jr. and G. W. Stewart, *Derivatives and perturbations of eigenvectors*, SIAM J. Numer. Anal., 25 (1988), pp. 679–691.

[27] J. K. REID, *A note on the stability of Gaussian elimination*, J. Inst. Math. Appl., 8 (1971), pp. 374–375.

[28] G. W. STEWART, *Error bounds for approximate invariant subspaces for close linear operators*, SIAM J. Numer. Anal., 8 (1971), pp. 796–808.

[29] G. W. STEWART AND G. ZHANG, *On a direct method for the solution of nearly uncoupled Markov chains*, Numer. Math., 59 (1991), pp. 1–11.

[30] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[31] R. S. VARGA AND D.-Y. CAI, *On the* LU *factorization of M-matrices*, Numer. Math., 38 (1981), pp. 179–192.

[32] J. H. WILKINSON, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.

[33] ———, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX. ANAL. APPL.
Vol. 14, No. 3, pp. 813–829, July 1993

015

CHOOSING THE INERTIAS FOR COMPLETIONS OF CERTAIN PARTIALLY SPECIFIED MATRICES*

JEROME DANCIS†

Abstract. This paper classifies the ranks and inertias of Hermitian completions of certain band matrices and other partially specified Hermitian matrices with chordal graphs and specified main diagonals. Some results are also presented on positivity- and negativity-preserving Hermitian completions. To complete partially specified Hermitian matrices with chordal graphs the inductive scheme presented by Grone, Johnson, Sa, and Wolkowicz [*Linear Algebra Appl.*, 58 (1984), pp. 109–124] is used. To complete Hermitian band matrices the inductive scheme presented by Dym and Gohberg [*Linear Algebra Appl.*, 36 (1981), pp. 1–24] is used. In both schemes, each inductive step is a *one-step completion* problem. At each inductive step, the classification of the kernels of one-step completions is used [*Linear Algebra Appl.*, 128 (1990), pp. 117–132]. This allows one to choose both the rank and the inertia of Hermitian completions of certain partial specified Hermitian matrices, while in Johnson and Rodman [*Linear and Multilinear Algebra*, 16 (1984), pp. 179–195], the rank cannot be chosen and the inertia can be chosen only for maximal rank completions.

Key words. matrix, Hermitian, rank, inertia, completion, minimal rank

AMS subject classification. 15A57

1. Introduction.

DEFINITION. We say that an $n \times n$ matrix $R = (r_{jk})$ is an *m-band* matrix with *bandwidth m* if $r_{jk} = 0$ for all $|k-j| > m$, and an $n \times n$ Hermitian matrix $F = (f_{jk})$ is a *completion* of such a matrix R if $f_{jk} = r_{jk}$ for all $|k-j| \leq m$. A *partial Hermitian matrix A* is one in which some entries are specified and others are left as free variables (and $a_{ij} = a_{ji}^*$ among the specified entries). A Hermitian matrix $F = (f_{jk})$ is a *completion* of such a matrix if $F = A$ at the specified entries.

DEFINITIONS. The *inertia* of a Hermitian matrix H is a triple In $H = (\pi, \nu, \delta)$ consisting of the numbers of positive, negative, and zero eigenvalues of H. We let $\pi(H), \nu(H)$, and $\delta(H)$ denote the three coordinates of In H. The $\pi(H)$, $\nu(H)$, and $\delta(H)$ are called the *positivity*, *negativity*, and *nullity* of H. The *positivity* of a partial specified Hermitian matrix A is the maximum number of positive eigenvalues for all the principal submatrices of A. The *negativity* of a partial specified Hermitian matrix A is the maximum number of negative eigenvalues for all the principal submatrices of A. The positivity and negativity of a partial specified Hermitian matrix A are denoted by $\pi^*(A)$ and $\nu^*(A)$. A Hermitian completion $F = (f_{jk})$ of a partial specified Hermitian matrix A is *positivity* (and *negativity*) *preserving* if $\pi(F) = \pi^*(A)$ (and $\nu(F) = \nu^*(A)$, respectively).

This paper presents our results on completing certain types of Hermitian band matrices and partial Hermitian matrices (with chordal graphs and specified main diagonals) to

(i) Hermitian matrices with prechosen inertia, or
(ii) Hermitian matrices with the same positivity and negativity.

In a companion paper [6] we will present our results on positive semidefinite completions. (See Theorem 1.13 at the end of this section.)

*Received by the editors March 5, 1990; accepted for publication (in revised form) October 17, 1991.

†University of Maryland, College Park, Maryland 20742-4015 (jnd@math.umd.edu).

These results are consequences of our classification in [4] of the kernels of "bordered" Hermitian matrices (defined in §2). This classification was a consequence of our generalized Poincaré inequalities.

DEFINITION. A *maximal submatrix* within a partial Hermitian matrix is a (Hermitian) principal submatrix that is maximal among the specified Hermitian principal submatrices. For an m-band $n \times n$ matrix R there are $n-m$ maximal Hermitian submatrices, each an $(m+1) \times (m+1)$ submatrix; we denote them by $R_1, R_2, \ldots, R_{n-m}$ when ordered from the upper left corner.

Dym and Gohberg [7] started this type of work by showing that, when all the maximal submatrices of a Hermitian band matrix R are positive definite, then R has a Hermitian completion F that is also positive definite, and F^{-1} is also a band matrix. They report that this result has connections with signal processing and system theory.

Grone, Johnson, Sa, and Wolkowicz [10] generalized the concept of a band matrix and a block-band matrix to that of a G-partial Hermitian matrix when G is a *chordal* graph. A *G-partial* Hermitian matrix is a partially specified Hermitian matrix whose associated undirected graph is G. This undirected graph G is *chordal* if it has no minimal simple circuit with four or more edges. (Equivalently, every polygonal simple closed curve with four or more edges has a chord.) The relevant properties of chordal graphs are listed early in §5. In [10], the chordal graph structure was used to organize the proof therein that there is a positive definite completion (positive semidefinite completion, respectively) of any G-partial Hermitian matrix R when G is a chordal graph and all the principal maximal submatrices of R are positive definite (positive semidefinite and the main diagonal of R is specified, respectively).

Of course, the inertia of a Hermitian completion F of a partial Hermitian $n \times n$ matrix A must be consistent with Poincaré's inequalities and Cauchy's interlacing theorem.

Poincaré's inequalities. For each principal submatrix K of a Hermitian matrix F,

$$\pi(F) \geq \pi(K) \text{ and } \nu(F) \geq \nu(K), \tag{1.1}$$

$$\pi(F) + \delta(F) \geq \pi(K) + \delta(K) \text{ and } \nu(F) + \delta(F) \geq \nu(K) + \delta(K). \tag{1.2}$$

We now state two special cases in which Poincaré's inequalities played an important role in the inertias of possible completions.

THEOREM 1.1 [11, Cor. 6]. *Let A be a partial Hermitian $n \times n$ matrix with a chordal graph whose maximal specified principal submatrices are all invertible. Given nonnegative integers π and ν with $\pi + \nu = n$, such that*

$$\pi \geq \pi^*(A) \text{ and } \nu \geq \nu^*(A),$$

there is an invertible *Hermitian completion F of A with* $\operatorname{In} F = (\pi, \nu, 0)$.

Johnson and Rodman's main result is the following.

THEOREM 1.2 [11]. *Each partial Hermitian matrix A with a chordal graph has a Hermitian completion F such that*

$$\nu(F) + \delta(F) = \operatorname{Max}\{\nu(K_i) + \delta(K_i)\},$$

and hence

$$\delta(F) \leq \operatorname{Max}\{\delta(K_i)\}$$

for the maximal specified principal submatrices $\{K_i\}$ of A.

Johnson and Rodman [11] leave open the question of choosing the ranks of Hermitian completions of partial and banded Hermitian matrices. They also leave open the question of choosing inertias for *singular* Hermitian completions of partial and banded Hermitian matrices even when all the specified principal matrices are invertible. This paper will address these two questions.

We have extended Poincaré's inequalities to Theorem 1.2 of [4] on bounded self-adjoint operators. We restate part of this theorem as Theorem 1.3 below. The symbols P, PS, π, π_1, ν, and ν_1 of Theorem 1.2 of [4] become $H, H_0, \pi(H), \pi(H_0), \nu(H)$, and $\nu(H_0)$, respectively, of Theorem 1.3 here; Δ remains as Δ.

THEOREM 1.3 (generalized Poincaré inequalities). *Let*

$$H = \begin{pmatrix} H_0 & A \\ A^* & B \end{pmatrix}$$

be a 2×2 *block Hermitian matrix with* B *an* $r \times r$ *submatrix. We set*

$$\Delta = \text{rank}(H_0, A)^* - \text{rank}(H_0) = \text{Dim Ker}H_0 - \text{Dim Ker}(H_0, A)^*.$$

Then (a) $\pi(H) \geq \pi(H_0) + \Delta$ *and* (b) $\nu(H) \geq \nu(H_0) + \Delta$.

Remark. The inertia of every Hermitian completion of a partial specified matrix must be consistent with Theorem 1.3 in addition to being consistent with Poincaré's inequalities.

The following *constrained* Hermitian completion result is a special case of Extension Theorem 1.3 of [2]. It is an example of being able to build a Hermitian completion (of a partial matrix) with any given inertia, which is consistent with Theorem 1.3. Here, in addition to being able to choose the inertia of the completion, we can choose the ranks of a column decomposition as well. The r_i and Δ_i of Theorem 1.4 below were the $(\pi_i + \nu_i + \delta_i - d_i)$ and $(\delta_i - d_i)$ of Extension Theorem 1.3 of [2].

THEOREM 1.4. *Given* s *Hermitian matrices* $H_{11}, H_{22}, \ldots, H_{ss}$, *with inertias* $\text{In}H_{ii} = (\pi_i, \nu_i, \delta_i)$, *let each* $n_i = \pi_i + \nu_i + \delta_i$ *and let* $n = \Sigma n_i$. *Let* S *be the block diagonal matrix with* $H_{11}, H_{22}, \ldots, H_{ss}$ *as the blocks on the main diagonal. Choose any nonnegative integers* $r_1, r_2, \ldots, r_s$ *such that*

$$\text{Rank } H_{ii} \leq r_i \leq n_i \quad \textit{for each } i = 1, 2, \ldots, s.$$

Set

$$\Delta_i = r_i - \text{Rank } H_{ii} \quad \textit{for each } i = 1, 2, \ldots, s.$$

Choose any nonnegative integers π, ν, *and* δ *such that* $n = \pi + \nu + \delta$ *and*

$$\pi \geq \text{Max}\{\pi_i + \Delta_i\}, \quad \nu \geq \text{Max}\{\nu_i + \Delta_i\}, \quad \textit{and} \quad \delta \geq \Sigma\delta_i - \Delta_i.$$

Then there is a block column Hermitian completion $H = (M_1, M_2, \ldots, M_s)$ *of* S, *with each* M_i, *an* $n \times n_i$ *non-Hermitian completion of* H_{ii} *such that* In $H = (\pi, \nu, \delta)$ *and, in addition, each* M_i *satisfies the constraint equation* Rank $M_i = r_i$.

Additional results on Hermitian completions of band matrices or partial Hermitian matrices appear in [5]–[11].

The next example demonstrates the importance of Theorem 1.3 when one is trying to determine the possible ranks of Hermitian completions of partial specified matrices.

Example 1.5. Let us examine this 2-band Hermitian matrix

$$S(a,b,c)=\begin{pmatrix}4&0&0&a^*&b^*\\0&0&0&1&c^*\\0&0&0&0&2\\a&1&0&0&0\\b&c&2&0&0\end{pmatrix}.$$

Here,

$$R_1=\begin{pmatrix}4&0&0\\0&0&0\\0&0&0\end{pmatrix},\quad R_2=\begin{pmatrix}0&0&1\\0&0&0\\1&0&0\end{pmatrix},\quad\text{and}\quad R_3=\begin{pmatrix}0&0&2\\0&0&0\\2&0&0\end{pmatrix},$$

$$\text{(1.3)}\qquad \text{In } R_1=(1,0,2)\quad\text{and}\quad \text{In } R_2=(1,1,1)=\text{In } R_3.$$

The submatrix $\left(\begin{smallmatrix}1&0\\c&2\end{smallmatrix}\right)$ ensures that

$$\text{(1.4)}\qquad [\text{Ker}\,(R_1\oplus I_2)]\cap\text{Ker}\,S(a,b,c)=\underline{0}$$

for all a, b, and c. Therefore (1.3), (1.4), and Theorem 1.3 with $\Delta=2$ imply that In $S(a,b,c)\geq(3,2,0)$. We note that $S(a,b,c)$ is a 5×5 matrix. Thus, consistency with Theorem 1.3 implies that

$$\text{In}\,S(a,b,c)=(3,2,0).$$

In contrast, the Poincaré inequalities alone only imply that

$$\text{In}\,S(a,b,c)\geq(1,1,0),$$

$$\pi(S(a,b,c))+\delta(S(a,b,c))\geq2,$$

and

$$\nu(S(a,b,c))+\delta(S(a,b,c))\geq2.$$

Hence, the Poincaré inequalities alone do not rule out various impossible inertia such as (1,1,3), (2,3,0), or (1,3,1), which are ruled out by Theorem 1.3.

In the last example, we saw the role Theorem 1.3 plays as it is applied directly to a partial Hermitian matrix. The next example is more subtle.

EXAMPLE 1.6. Let

$$F(z)=\begin{pmatrix}0&z^*\\z&0\end{pmatrix}.$$

Then Rank $F(0)=0$ and Rank $F(z)=2$ for all $z\neq0$.

Remark. In this example, it is possible to obtain both low-rank and high-rank completions. It is *not* possible to obtain an intermediary-rank completion, that is, Rank $F(z)$ is never one. Here, when one introduces a positive eigenvalue, the kernel of a maximal specified submatrix will not appear in the kernel of the Hermitian completion, and this (by Theorem 1.3) forces the existence of a (possibly unwanted) negative eigenvalue as well.

These last two examples and Example 1.12 as well demonstrate that it is *not* always possible to build Hermitian completions with all the inertias that are consistent with Poincaré's inequalities.

In Theorems 1.7 and 1.11, we present situations in which Hermitian completions that have all the possible inertias consistent with Poincaré's inequalities *may* be constructed.

THEOREM 1.7. *Given a Hermitian m-band $n \times n$ matrix R, suppose that each of the $n-m$ maximal submatrices $R_1, R_2, \ldots, R_{n-m}$ of R is an invertible matrix. For any potential inertia $(\pi, \nu, \delta), \pi + \nu + \delta = n$, such that*

$$\pi \geq \operatorname{Max}\{\pi(R_i), i = 1, 2, \ldots, n-m\} \quad \textit{and} \quad \nu \geq \operatorname{Max}\{\nu(R_i), i = 1, 2, \ldots, n-m\},$$

there is a Hermitian completion F of R such that

$$\operatorname{In} F = (\pi, \nu, \delta).$$

The proof of this theorem is presented in §4.

As noted in Theorem 1.1, the case when the Hermitian completion F of Theorem 1.7 is invertible (and hence $\delta = 0$), was established by Johnson and Rodman [11].

DEFINITION. The maximal Hermitian $m \times m$ submatrices of the $m-1$-band $n \times n$ matrix made from an m-band matrix R by setting the outside pair of diagonals to zeros shall be called the *almost-maximal Hermitian submatrices.* We denote these $n-m+1$ submatrices by $R_1^*, R_2^*, \ldots, R_{n-m+1}^*$ when ordered from the upper-left corner.

In our earlier paper [4], we found it useful to describe the connections between the kernels of submatrices of a given matrix in the following manner.

Notation for identified subspaces. Given submatrices

$$S_k = \{a_{ij}, L_k \leq i \leq N_k \text{ and } J_k \leq j \leq K_k\}$$

of an $n \times m$ matrix M, there are the natural injection maps

$$\iota_k : \text{Domain of } S_k \to \text{ Domain of } M = \mathbf{C}^m,$$

where $\iota_k(v) = (0, 0, \ldots, 0, v, 0, \ldots, 0)$ and there are $J_k - 1$ zeros placed before the vector v and $m - K_k - 1$ zeros placed after the v, for each k. Thus each $\iota_k(\operatorname{Ker} S_k) \subset \mathbf{C}^m$, and we identify $\operatorname{Ker} S_k$ with $\iota_k(\operatorname{Ker} S_k)$, for each k. More generally, for any submatrix S of M, there is the orthoprojection

$$P : \text{Domain of } M \to \text{ Domain of } S$$

and the natural (coordinate by coordinate) injection map

$$\iota : \text{Domain of } S \to \text{ Domain of } M = \mathbf{C}^m$$

such that P is the right inverse of ι. We identify each $\operatorname{Ker} S$ with $\iota(\operatorname{Ker} S)$. Therefore, we will write

$$\operatorname{Ker} S_1 + \operatorname{Ker} S_2 + \cdots + \operatorname{Ker} S_r$$

as a shorthand notation for the nondirect sum of the subspaces

$$\iota_1(\operatorname{Ker} S_1) + \iota_2(\operatorname{Ker} S_2) + \cdots + \iota_s(\operatorname{Ker} S_r).$$

We now present our main results on positivity- and negativity-preserving Hermitian completions as Theorems 1.8 and 1.10 below.

THEOREM 1.8 (positivity- and negativity-preserving completions). *Given a Hermitian m-band matrix R, let $R_1, R_2, \dots, R_{n-m}$ and $R_1^*, R_2^*, \dots, R_{n-m+1}^*$ be the maximal and almost-maximal Hermitian submatrices of R (all ordered from the upper-left corner). Suppose that*

$$(1.5) \qquad \delta(R_i) \geq \delta(R_i^*) \quad \text{and} \quad \delta(R_{i+1}) \geq \delta(R_i^*) \quad \text{for each} \quad i = 1, 2, \dots, n-m.$$

Then there is a Hermitian completion F of R, such that

$$\pi(F) = \operatorname{Max}\{\pi(R_i), i = 1, 2, \dots, n-m\} \text{ and } \nu(F) = \operatorname{Max}\{\nu(R_i), i = 1, 2, \dots, n-m\}.$$

Also,

$$\operatorname{Ker} F \supset \operatorname{Ker} R_1 + \operatorname{Ker} R_2 + \cdots + \operatorname{Ker} R_{n-m},$$

where each Ker R_i *is considered as a subspace of* $\mathbf{C}^n$ *in the manner of the notation for identified subspaces.*

The proof of this theorem is presented in §3.

Remark. We note that the band matrix in Example 1.5 shows that the conclusions of Theorem 1.8 need *not* be achieved when the hypotheses are not satisfied.

The next example shows that a minimal rank Hermitian completion may be unique even when the minimal rank non-Hermitian completions are plentiful.

EXAMPLE 1.9. Let R be the 0-band matrix

$$R = \begin{pmatrix} 0 & ? & ? \\ ? & 0 & ? \\ ? & ? & 1 \end{pmatrix}.$$

We observe that R has an infinite number of completions with rank one, but that

$$F = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is the *unique* minimal rank *Hermitian* completion with rank one.

THEOREM 1.10 (positivity- and negativity-preserving completions). *Let R be a G-partial Hermitian $n \times n$ matrix where G is a chordal graph and the main diagonal of R is specified. Let $R_1, R_2, \dots, R_s$ be the maximal specified principal submatrices of R. For each principal submatrix S_j of a specified maximal principal submatrix R_i of R, we suppose that*

$$\operatorname{Ker} R_i \supset \operatorname{Ker} S_j.$$

Then there is a Hermitian completion F of R, such that

$$\pi(F) = \operatorname{Max}\{\pi(R_i), i = 1, 2, \dots, s\} \text{ and } \nu(F) = \operatorname{Max}\{\nu(R_i), i = 1, 2, \dots, s\}.$$

Also,

$$\operatorname{Ker} F \supset \operatorname{Ker} R_1 + \operatorname{Ker} R_2 + \cdots + \operatorname{Ker} R_s$$

(*where each* Ker R_i *is considered as a subspace of* $\mathbf{C}^n$ *in the manner of the notation for identified subspaces*).

Remark. Because of Poincaré's inequalities (1.1), we observe that the matrix F supplied by Theorems 1.8 and 1.10 must have the *minimum* possible rank among all Hermitian completions of the given matrix R.

The proofs of Theorems 1.10 and 1.11 are presented in §5.

We admit that the condition $\operatorname{Ker} R_i \supset \operatorname{Ker} S_j$ in Theorem 1.10 is quite strong and unnatural. Nevertheless, Theorem 1.10 is crucial for our proof of Theorem 1.11.

Remark. For band matrices we only required that the *maximal* principal submatrices be invertible (in Theorem 1.7), but for partial matrices with chordal graphs our proof will require that all (including the nonmaximal) principal submatrices be invertible in order that we may freely choose the inertia (Theorem 1.11).

THEOREM 1.11. *Let R be a G-partial Hermitian $n \times n$ matrix where G is a chordal graph and the main diagonal of R is specified. Suppose that all the specified principal submatrices $R_1, R_2, \dots, R_s$ of R are invertible. For all nonnegative integers $(\pi, \nu, \delta), \pi + \nu + \delta = n$, such that*

$$\pi \geq \operatorname{Max}\{\pi(R_i), i = 1, 2, \dots, s\} \quad \text{and} \quad \nu \geq \operatorname{Max}\{\nu(R_i), i = 1, 2, \dots, s\},$$

there is a Hermitian completion F of R such that

$$\operatorname{In} F = (\pi, \nu, \delta).$$

Remark. All the results stated in this paper are also valid for real symmetric completions when the partial Hermitian $n \times n$ matrix is a real symmetric matrix.

EXAMPLE 1.12. Let

$$R = \begin{pmatrix} 2 & 2 & ? \\ 2 & x & 1 \\ ? & 1 & 1 \end{pmatrix},$$

where $x = R(2,2)$ is unspecified.

We note that the graph is chordal and that each of the two specified principal submatrices of R is invertible, and has rank one and inertia (1,0,0). But as soon as x is specified there will be a 2×2 specified principal submatrix with rank two and inertia not equal to (1,0,1) (since x cannot be both 1 and 2). Hence, there is no Hermitian completion with inertia (1,0,2).

Remark. This matrix R satisfies all the hypotheses of Theorems 1.10 and 1.11 except that the main diagonal is not specified and there is no positivity- and negativity-preserving Hermitian completion of R with inertia (1,0,2). Thus Example 1.12 demonstrates the necessity of the condition that the main diagonal be specified in the statement of these two theorems.

Remark. Here we announce the next theorem, which shows that there are no gaps in the possible ranks for the positive semidefinite Hermitian completions of a partial Hermitian matrix with a chordal graph and a specified main diagonal. This is in contrast to the situation in Example 1.6 and in Lemma 2.4 (when $\operatorname{In} H_1 = \operatorname{In} H_3$), where there is a gap in the possible ranks for the (one-step) Hermitian completions of the band matrices listed there.

THEOREM 1.13 (possible ranks for positive semidefinite completions). *Let R be a G-partial Hermitian $n \times n$ matrix where G is a chordal graph and the main diagonal of R is specified. Suppose that each of the maximal submatrices $R_1, R_2, \dots, R_s$ is a positive semidefinite matrix. Then there is a positive semidefinite completion F of R with* Rank $F = \pi$ *if and only if*

$$\pi \geq \operatorname{Max}\{\operatorname{Rank} R_i, i = 1, 2, \dots, s\}$$

and

$$\pi \leq n - \operatorname{Dim}(\operatorname{Ker} R_1 + \operatorname{Ker} R_2 + \dots + \operatorname{Ker} R_s).$$

This theorem will be established in a companion paper [6].

2. Kernels of bordered matrices. The basis for this and many papers on completing band matrices is an examination of the situation for bordered matrices.

Bordered matrix hypotheses. Let H_2 be an $(r-2) \times (r-2)$ Hermitian matrix. Let v and w be vectors in C^{r-2} and let a and b be real numbers. Set

$$H_1 = \begin{pmatrix} a & v^* \\ v & H_2 \end{pmatrix} \quad \text{and} \quad H_3 = \begin{pmatrix} H_2 & w \\ w^* & b \end{pmatrix}$$

and let $H(z)$ be the bordered matrix

$$H(z) = \begin{pmatrix} a & v^* & z* \\ v & H_2 & w \\ z & w^* & b \end{pmatrix}.$$

DEFINITION. $H(z)$ is called a *one-step completion* of $H(0)$.

Remark. In [4], we classified the kernels of bordered matrices. Using Claim 3.10 of [4], one may easily check that the seven Lemmas 3.3–3.9 of [4] cover all the possible cases for bordered matrices. Therefore, we are able to use this classification together with the usual method for constructing Hermitian completions of band matrices in order to establish a variety of results about Hermitian completion of band matrices. In fact, this is exactly what we did both in this paper and in our other papers [5], [6]. Ellis and Lay use a similar approach in [9].

The lemmas in this section are largely restatements of lemmas in [4], together with small extensions. The theorems in this section are largely common threads of these lemmas. These theorems, in turn, will be used in each inductive step in the proofs of the theorems of §1.

The next result, which was a simple consequence of the generalized Poincaré inequalities (Theorem 1.3) was established as Corollary 3.2 on semi-bordered matrices in [4].

LEMMA 2.1 (semi-bordered matrices). *Let H_1 and H_2 be as in the bordered matrix hypotheses.*

(a) *Then,* $|\delta(H_1) - \delta(H_2)| \leq 1$,

(b) $\delta(H_1) = \delta(H_2) \leftrightarrow \operatorname{Ker} H_1 = \operatorname{Ker}\ (1 \oplus H_2)$,

(c) $\delta(H_1) = \delta(H_2) + 1 \leftrightarrow \operatorname{Ker} H_1 \supset \operatorname{Ker} (1 \oplus H_2) \leftrightarrow \operatorname{In} H_1 = \operatorname{In} H_2 + (0,0,1)$,

(d) $\delta(H_1) = \delta(H_2) - 1 \leftrightarrow \operatorname{Ker} (1 \oplus H_2) \supset \operatorname{Ker} H_1 \leftrightarrow \operatorname{In} H_1 = \operatorname{In} H_2 + (1,1,-1)$.

Thus the connection between the dimensions of the kernels determines the connection between the kernels. We will use this to show, in the lemmas and theorems of this section, that the connections between $\delta(H_1), \delta(H_2)$, and $\delta(H_3)$ almost determine $\delta(H(z))$.

OBSERVATION 2.2. All the results of Lemma 2.1 about $\{H_1, H_2\}$ are applicable to the ordered pair of matrices $\{H(z), H_1\}$, also to each pair $\{H(z), H_3\}$ and $\{H_3, H_2\}$ with the obvious modifications like $\operatorname{Ker} H_3 = \operatorname{Ker} (H_2 \oplus 1)$ in (b).

We will implicitly use this observation often in the proofs in this section. We do this when we establish the following corollary to Lemma 2.1 on semi-bordered matrices.

COROLLARY 2.3. *Let the submatrices H_1, H_2, and H_3 overlap in $H(z)$ in the manner described in the bordered matrix hypotheses. Suppose, for some number z_0, that*

$$\delta(H(z_0)) \geq \delta(H_1) \quad \textit{and} \quad \delta(H(z_0)) \geq \delta(H_3),$$

then

$$\operatorname{Ker} H(z_0) \supset \operatorname{Ker} (H_1 \oplus 1) + \operatorname{Ker} (1 \oplus H_3). \tag{2.1}$$

LEMMA 2.4. *Let the submatrices H_1, H_2, and H_3 overlap in $H(z)$ in the manner described in the bordered matrix hypotheses. Suppose that*

$$\delta(H_1) = \delta(H_2) + 1 \quad \textit{and} \quad \delta(H_3) = \delta(H_2) \quad \textit{or} \quad \delta(H_1)$$

(*or, equivalently,* Rank H_1 = Rank H_2 *and* Rank $H_3 \leq 1 +$ Rank H_1).

Then there is a unique number z_0 such that

(i) $\operatorname{In} H(z) = \operatorname{In} H_1 + (1, 1, -1)$ *for all* $z \neq z_0$, *and*

(ii) $\operatorname{In} H(z_0) = \operatorname{In} H_3 + (0, 0, 1)$; *also* (2.1) *is satisfied.*

Furthermore, when H_1 and H_3 are real matrices, then the number z_0 is a real number.

Remark. The hypotheses of this lemma are equivalent to the hypotheses of Lemma 3.3 of [4]. In [4], we presented our results in terms of the nullities $\delta_i = \delta(H_i), i = 1, 2,$ and 3, and $\delta(z_0) = \delta(H(z_0))$ of the matrices that make up the bordered matrix. There is a related result in [9].

Proof. The number z_0 is supplied by Lemma 3.3(b) of [4]. That z_0 is a real number when H_1 and H_3 are real matrices follows from the formula defining z_0 in the proof of Lemma 3.3 of [4].

Lemma 2.4(i) and $\operatorname{In} H(z_0) = \operatorname{In} H_3 + (0, 0, 1)$ of part (ii) were established in Lemma 3.3 of [4]. Thus, only (2.1) remains to be proven. The equation $\operatorname{In} H(z_0) = \operatorname{In} H_3 + (0, 0, 1)$ implies that $\delta(H(z_0)) = 1 + \delta(H_3)$. This and Rank $H_3 \leq 1 +$ Rank H_1 implies that $\delta(H(z_0)) \geq \delta(H_1)$. Corollary 2.3, together with these conditions on the nullities, implies (2.1). □

We restate Lemma 3.5 of [4] in the following way.

LEMMA 2.5. *Let the submatrices H_1, H_2, and H_3 overlap in $H(z)$ in the manner described in the bordered matrix hypotheses. Suppose that*

$$\delta(H_1) = \delta(H_2) = \delta(H_3) \quad \textit{and} \quad \nu(H_1) = \nu(H_2) + 1 = \nu(H_3) + 1.$$

Then

$$\operatorname{In} H(z) = \operatorname{In} H_1 + (1, 0, 0) = \operatorname{In} H_2 + (1, 1, 0) + \operatorname{In} H_3 + (0, 1, 0),$$

and

$$\operatorname{Ker} H(z) = \operatorname{Ker}(H_1 \oplus 1) = \operatorname{Ker}(1 \oplus H_2 \oplus 1) = \operatorname{Ker}(1 \oplus H_3),$$

for all z.

LEMMA 2.6 [4], [8]. *Let the submatrices H_1, H_2, and H_3 overlap in $H(z)$ in the manner described in the bordered matrix hypotheses. Suppose that*

$$\delta(H_1) = \delta(H_2) = \delta(H_3) \quad \textit{and} \quad \operatorname{In} H_1 = \operatorname{In} H_3;$$

then there is a circle C in the complex plane such that

(a) $\operatorname{In} H(z) = \operatorname{In} H_1 + (1, 0, 0)$, $\operatorname{Ker} H(z) = \operatorname{Ker}(H_1 \oplus 1) = \operatorname{Ker}(1 \oplus H_2 \oplus 1) = \operatorname{Ker}(1 \oplus H_3)$ *for all complex numbers z outside of this circle C*;

(b) $\operatorname{In} H(z) = \operatorname{In} H_1 + (0, 1, 0)$, $\operatorname{Ker} H(z) = \operatorname{Ker}(H_1 \oplus 1) = \operatorname{Ker}(1 \oplus H_2 \oplus 1) = \operatorname{Ker}(1 \oplus H_3)$ *for all complex numbers z inside of this circle C*;

(c) *Equation* (2.1) *holds, and* $\operatorname{In} H(z) - (0, 0, 1) = \operatorname{In} H_1 = \operatorname{In} H_3$, *for all complex numbers z on this circle C.*

Remark. For real symmetric matrices, there are real numbers x_1, x_2, and x_0, such that the three cases of this lemma are satisfied; that is, $\operatorname{In} H(x_1) = \operatorname{In} H_1 + (1, 0, 0)$, $\operatorname{In} H(x_2) = \operatorname{In} H_1 + (0, 1, 0)$, and $\operatorname{In} H(x_0) = \operatorname{In} H_1 + (0, 0, 1)$.

Proof. Lemma 2.6 is largely a restatement of Lemma 3.6 of [4], which in turn was a consequence of Theorem 1.1 of [8]; the new result is that (2.1) holds in part (c), and this follows from Corollary 2.3 and the inertia equation in part (c).

Proof of Remark. The circle of Lemma 2.6 is supplied by the statement of Theorem 1.1 of [8]. For real symmetric matrices, (1.1) of [8] implies that the center of the circle is a real number. This implies that the real numbers x_1, x_2, and x_0 exist and satisfy the inertial equations listed in the remark. □

Remark 2.7. For a bordered matrix, $H(z)$, all the cases when $\delta(H_1) = \delta(H_2) = \delta(H_3)$ are covered by Lemma 2.6 or by Lemma 2.5, either by itself or by interchanging H_1 and H_3.

Proof. Let the submatrices H_1, H_2, and H_3 overlap in $H(z)$ in the manner described in the bordered matrix hypotheses. We note that Poincaré's inequalities imply that $\nu(H_1) - \nu(H_2) \epsilon \{0,1\}$ and $\nu(H_3) - \nu(H_2) \epsilon \{0,1\}$, which together imply that $|\nu(H_1) - \nu(H_3)| \leq 1$. Therefore, there are only these three possibilities:

(i) $\nu(H_1) - \nu(H_3) = 1$, which is the case of Lemma 2.5;

(ii) $\nu(H_1) - \nu(H_3) = 0$, which implies that In H_1 = In H_3, hence this is the case of Lemma 2.6;

(iii) $\nu(H_1) - \nu(H_3) = -1$, which is the case of Lemma 2.5 with H_1 and H_3 interchanged.

THEOREM 2.8. *Let the submatrices H_1, H_2, and H_3 overlap in $H(z)$ in the manner described in the bordered matrix hypotheses. If*

$$\delta(H_2) \leq \delta(H_1) \quad \textit{and} \quad \delta(H_2) \leq \delta(H_3) \tag{2.2}$$

(*or, equivalently, if* $2 + \text{Rank } H_2 \neq \text{Rank } H_1$ *or* Rank H_3), *then there is a number* z_0, *such that* (2.1) *holds and*

$$\pi(z_0) = \text{Max}\{\pi(H_1), \pi(H_3)\} \quad \textit{and} \quad \nu(z_0) = \text{Max}\{\nu(H_1), \nu(H_3)\}, \tag{2.3}$$

and

$$\delta(H_1) \leq \delta(H(z_0)) \quad \textit{and} \quad \delta(H_3) \leq \delta(H(z_0)). \tag{2.4}$$

Also, when H_1 and H_3 are real matrices, then the number z_0 may be chosen to be a real number.

Remark. This theorem is the common thread of Lemmas 2.4–2.6; with the exception of (2.1), this theorem was established as Theorem 1.2 in [4].

LEMMA 2.9. *Let vectors v and w and submatrices H_1, H_2, and H_3 overlap in $H(z)$ in the manner described in the bordered matrix hypotheses. Suppose that both v and w miss the image space $H_2(C^{r-2})$, and that the* span $\{v, w\}$ *meets* $H_2(C^{r-2})$ *in a line. Then*

$$\text{In}\, H_1 = \text{In}\, H_2 + (1,1,-1) = \text{In}\, H_3 \quad \textit{and} \quad \text{Ker}\,(H_1 \oplus 1) = \text{Ker}\,(1 \oplus H_3)$$

and (2.1) *holds for all z.*

Also, there are two open half-planes P_+ and P_- whose closures meet in a line **l**, *such that:*

(a) $\text{In}\, H(z) = \text{In}\, H_1 + (1,0,0) = \text{In}\, H_2 + (2,1,-1)$, *and* $\text{Ker}\, H(z) = \text{Ker}\,(H_1 \oplus 1) = \text{Ker}\,(1 \oplus H_3)$ *for all complex numbers z in P_+;*

(b) $\operatorname{In} H(z) = \operatorname{In} H_1 + (0,1,0)$, *and* $\operatorname{Ker} H(z) = \operatorname{Ker}(H_1 \oplus 1) = \operatorname{Ker}(1 \oplus H_3)$ *for all complex numbers* z *in* P_-;

(c) $\operatorname{In} H(z) - (0,0,1) = \operatorname{In} H_1 = \operatorname{In} H_3$ *for all complex numbers* z *on the line* **l**.

Furthermore, when $H(0)$ *is a real matrix, then this line* **l** *will meet the real axis at a unique real number.*

Proof. The hypotheses of Lemma 3.9 of [4] are the same as the hypotheses of Lemma 2.9 here. The hypotheses are symmetric with respect to H_1 and H_3. Therefore, Lemma 3.9 of [4] itself and again with H_1 and H_3 interchanged will establish this lemma except for (2.1) in part (c). Corollary 2.3 here and Lemma 3.9 of [4] will establish (2.1) in part (c). □

THEOREM 2.10. *Let the submatrices* H_1, H_2, *and* H_3 *overlap in* $H(z)$ *in the manner described in the bordered matrix hypotheses. Suppose that* $\operatorname{Ker}(H_1 \oplus 1) = \operatorname{Ker}(1 \oplus H_3)$. *Then there is a number* z, *such that* (2.1) *holds. Also, either case* (a) *or* (b) *exists*:

(a) $\operatorname{In} H_1 \neq \operatorname{In} H_3$. *In this case,* $\operatorname{In} H_1 - \operatorname{In} H_3 = \pm(1,-1,0)$ *and* $\nu(H(z)) = \operatorname{Max}\{\nu(H_1), \nu(H_3)\}$ *and* $\pi(H(z)) = \operatorname{Max}\{\pi(H_1), \pi(H_3)\}$ *for all complex numbers* z; *or*

(b) $\operatorname{In} H_1 = \operatorname{In} H_3$. *In this case, there are complex numbers* z_0, z_1, *and* z_2, *such that* (2.1) *is valid, and* $\operatorname{In} H(z_0) - (0,0,1) = \operatorname{In} H_1 = \operatorname{In} H_3 = \operatorname{In} H(z_1) - (1,0,0) = \operatorname{In} H(z_2) - (0,1,0)$. *If* H_1 *and* H_3 *are real matrices then* z_0, z_1, *and* z_2 *may be chosen as real numbers.*

Proof. Using Claim 3.10 of [4], one sees that Lemmas 3.3–3.9 of [4] cover all the possible cases for bordered matrices.

Claim (i). The hypotheses of this theorem are not consistent with Lemmas 3.4 and 3.8 of [4].

Proof. Because of the overlapping of the matrices H_1 and H_3 in the bordered matrix $H(z)$, the condition $\operatorname{Ker}(H_1 \oplus 1) = \operatorname{Ker}(1 \oplus H_3)$ implies that

$$\operatorname{Ker}(1 \oplus H_2) \supset \operatorname{Ker} H_1 \quad \text{and} \quad \operatorname{Ker}(H_2 \oplus 1) \supset \operatorname{Ker} H_3.$$

Now these conditions are not consistent with the hypotheses of Lemmas 3.4 and 3.8 of [4].

Claim (ii). The hypotheses of this theorem are not consistent with Lemma 3.3 of [4].

Proof. The hypothesis of Lemma 3.3 of [4] is the same as the hypothesis of Lemma 2.4 here. They say that $\delta(H_1) = \delta(H_2) + 1$, which implies that there is a vector in $\operatorname{Ker} H_1$ with a *nonzero* first coordinate. This is not consistent with the equation $\operatorname{Ker}(H_1 \oplus 1) = \operatorname{Ker}(1 \oplus H_3)$ in this theorem. □

Claim (iii). The hypotheses of this theorem are not consistent with Lemma 3.7 of [4].

Proof. We assume the bordered matrix hypotheses. The hypothesis of Lemma 3.7 of [4] states that the pair of vectors $\{v, w\}$ (in the bordered matrix) is a linearly independent set whose span meets the image space $H_2(C^{r-2})$ only at $\underline{0}$. This is not consistent with the equation $\operatorname{Ker}(H_1 \oplus 1) = \operatorname{Ker}(1 \oplus H_3)$ in this theorem. □

We have eliminated all but Lemmas 3.5, 3.6, and 3.9 of [4] from the classifying list of Lemmas 3.3–3.9 of [4], which, as noted, cover all the possible cases for bordered matrices. Therefore, by the process of elimination, the hypotheses of this theorem must be included in the hypotheses of Lemmas 3.5, 3.6, and 3.9 of [4], which have been restated here as the hypotheses of Lemmas 2.5, 2.6, and 2.9. Hence, part (a) of this theorem is determined by Lemma 2.5, and part (b) is determined by Lemmas 2.6 and 2.9. □

3. Positivity- and negativity-preserving completions. The problem of this section is to find conditions on a band matrix R that will enable us to build positivity- and negativity-preserving Hermitian completion of R.

EXAMPLE 3.1. Let

$$H(z) = \begin{pmatrix} 7 & 0 & z \\ 0 & 0 & 0 \\ z^* & 0 & -7 \end{pmatrix} \quad \text{and} \quad G(z) = \begin{pmatrix} 7 & 0 & z \\ 0 & 0 & 0 \\ z^* & 0 & 7 \end{pmatrix}.$$

Then Rank $H(z) = 2$ for all z, and Rank $G(z) = 2$ for all $z \neq 7$, but Rank $G(7) = 1$. Thus the minimal rank of $H(z)$ is 2, and the minimal rank of $G(z)$ is 1 even though the ranks of any pair of corresponding submatrices of $H(z)$ and $G(z)$ (without z and z^*) are equal.

Remark. This example demonstrates that the minimal rank of a Hermitian completion of a bordered Hermitian matrix is not always a function of just the ranks of its submatrices as it is for the non-Hermitian completions of triangular matrices [1].

DEFINITION. A *simple diagonal completion* R' of an m-band Hermitian matrix R is an $m+1$-band Hermitian completion of R. An Nth (*successive*) *simple diagonal completion* of an m-band Hermitian matrix R is a simple diagonal completion of an $(N-1)$st (successive) simple diagonal completion of R. (The original matrix R is the zeroth simple diagonal completion of itself.)

OBSERVATION 3.2 (Dym and Gohberg [7]). For a band matrix the two maximal and the almost maximal submatrices R_i, R_{i+1}, and R_i^* always overlap precisely as the H_1, H_3, and H_2 of the bordered matrix hypotheses.

The standard method for constructing Hermitian completions of band matrices. Construct successive simple diagonal completions; namely, complete R to R', then complete R' to $R'' = (R')', \ldots$, to F, where F is the $(n-m+1)$st simple diagonal completion of R. Using Observation 3.2, each successive simple diagonal completion is achieved by a set of independent one-step completions.

This method was pioneered by Dym and Gohberg in [7]; it was also used in [5], [8], and [9].

Theorem 1.8 will be established largely as a consequence of the standard method for constructing Hermitian completions of band matrices using Theorem 2.8 at each one-step completion.

Proof of Theorem 1.8 *on positivity- and negativity-preserving completions.* Let $R_1, R_2, \ldots, R_{n-m}$ and $R_1^*, R_2^*, \ldots, R_{n-m+1}^*$ be the maximal and almost-maximal Hermitian submatrices (all ordered from the upper-left corner) of a Hermitian m-band matrix R. Then,

$$\pi^*(R) = \operatorname{Max}\{\pi(R_i), i = 1, 2, \ldots, n-m\} \quad \text{and}$$
$$\nu * (R) = \operatorname{Max}\{\nu(R_i), i = 1, 2, \ldots, n-m\}.$$

Set

$$K = \operatorname{Ker} R_1 + \operatorname{Ker} R_2 + \cdots + \operatorname{Ker} R_{n-m}.$$

Using Observation 3.2, we will complete R to a simple diagonal completion R' by a series of $n-m-1$ one-step completions $H^j(z_j), j = 1, 2, \ldots, n-m-1$ (numbered from the upper-left corner).

We let H_1^j, H_2^j, and H_3^j denote the overlapping submatrices H_1, H_2, and H_3 of the jth bordered matrix $H^j(z_j)$. With this notation, we note that

$$H_1^j = R_j \quad \text{and} \quad H_3^j = R_{j+1} = H_1^{j+1} \quad \text{and} \quad H_3^{j+1} = R_{j+2}.$$

Also,

$$H_2^j = R^*_{j+1} \text{ and } H_2^{j+1} = R^*_{j+2}.$$

Therefore, (1.5) says that (2.2) is satisfied for these bordered matrices, and hence we may and do use Theorem 2.8 as we do each one-step completion as we complete R to a simple diagonal completion R'. Let $R'_i, i = 1, 2, \dots, n-m-1$, be the maximal submatrices of this simple diagonal completion R' (the R'_i are ordered from the upper-left corner). Now Theorem 2.8 provides (2.4), which for the jth one-step completion $H^j(z_j) = R'_j$, states that

$$\delta(R_j) \le \delta(R'_j) \text{ and } \underline{\delta(R_{j+1}) \le \delta(R'_j)}, \tag{3.1}$$

and for the $j+1$st one-step completion $H^{j+1}(z_{j+1})$, it states that

$$\underline{\delta(R_{j+1}) \le \delta(R'_{j+1})} \text{ and } \delta(R_{j+2}) \le \delta(R'_{j+1}), \tag{3.2}$$

Using Observation 3.2, we will complete R' to a simple diagonal completion R'' by a series of $(n-m-2)$ one-step completions $H'^j(z_j), j = 1, 2, \dots, n-m-2$ (numbered from the upper-left corner).

We let $H_1'^j$, $H_2'^j$, and $H_3'^j$ denote the overlapping submatrices H_1, H_2, and H_3 of the jth bordered matrix $H'^j(z_j)$. With this notation, we note that

$$H_1'^j = R'_j \text{ and } R'_{j+1} = H_3'^j \text{ and } H_2'^j = R_{j+1}.$$

Therefore, the *two underlined* nullity inequalities in (3.1) and (3.2) are condition (2.2) for the jth bordered matrix $H'^j(z_j)$ and condition (1.5) for the simple diagonal completion R'. Thus Theorem 2.8 is used in this manner to propagate condition (1.5) for a band matrix R to condition (1.5) for a simple diagonal completion R' of R.

Also, (2.3) will imply that

$$\pi^*(R') = \pi^*(R) \text{ and } \nu^*(R') = \nu^*(R).$$

Now we simply do induction on the bandwidth. This will result in

$$\pi(F) = \pi^*(F) = \pi^*(R) \text{ and } \nu(F) = \nu^*(F) = \nu^*(R).$$

Also, (2.4) and Observation 2.3 will result in

$$\operatorname{Ker} S_1^{(k)} + \operatorname{Ker} S_2^{(k)} + \cdots + \operatorname{Ker} S_{(n-m-k)}^{(k)} \supset K,$$

where the $S_j^{(k)}$ are the maximal submatrices of the kth simple diagonal completion of R. Hence, by induction, $\operatorname{Ker} F \supset K$, as desired. □

COROLLARY 3.3. *Given a Hermitian m-band matrix R. Suppose that each of the* $(n-m+1)$ almost-*maximal submatrices* $R_1^*, R_2^*, \dots, R_{n-m-1}^*$ (*when ordered from the upper-left corner*) *of R is an* invertible *matrix. Let* $R_1, R_2, \dots, R_{n-m}$ *be the maximal submatrices of R. Then there is a Hermitian completion F of R such that*

$$\pi(F) = \operatorname{Max}\{\pi(R_i), i = 1, 2, \dots, n-m\} \text{ and}$$
$$\nu(F) = \operatorname{Max}\{\nu(R_i), i = 1, 2, \dots, n-m\}.$$

Also,

$$\operatorname{Ker} F \supset \operatorname{Ker} R_1 + \operatorname{Ker} R_2 + \cdots + \operatorname{Ker} R_{n-m}.$$

Proof. Since the almost-maximal submatrices of R are invertible matrices, they have no kernels and therefore the dimension of the kernels cannot decrease when going from an almost-maximal submatrix to a maximal submatrix. Thus this corollary is a special case of Theorem 1.8.

LEMMA 3.4. *Given a Hermitian m-band matrix R, suppose that each of the $n-m$ maximal submatrices $R_1, R_2, \ldots, R_{n-m}$ of R is an invertible matrix. Then there is a Hermitian completion F of H, such that*

$$\pi(F) = \text{Max}\{\pi(R_i), i = 1, 2, \ldots, n-m\} \quad \textit{and}$$
$$\nu(F) = \text{Max}\{\nu(R_i), i = 1, 2, \ldots, n-m\}.$$

Proof. When going from the invertible maximal submatrices of R to the simple diagonal completion R', Theorem 2.10 is applicable. Therefore, we can construct R' without increasing the maximum numbers of positive or negative eigenvalues, and R' must satisfy the hypotheses of Corollary 3.3. Therefore, we can use Corollary 3.3 to complete R' (and R) to the desired matrix F. □

We now combine the ideas of Theorems 1.8 and 2.10.

THEOREM 3.5. *Given a Hermitian m-band matrix R. Let $R_1, R_2, \ldots, R_{n-m}$, and $R_1^*, R_2^*, \ldots, R_{n-m+1}^*$ be the maximal and almost-maximal Hermitian submatrices of R (all ordered from the upper-left corner). Suppose, for each $i = 1, 2, \ldots, n-m-1$, that one of these conditions is satisfied:*

(a) $\delta(R_i) \geq \delta(R_i^*)$ *and* $\delta(R_{i+1}) \geq \delta(R_i^*)$, *or*

(b) $\text{Ker}\,(R_i \oplus 1) = \text{Ker}\,(1 \oplus R_{i+1})$.

Then there is a simple diagonal completion R' of R that satisfies the hypotheses of Theorem 1.8, *and there is a Hermitian completion F of R' and R, such that*

$$\pi(F) = \text{Max}\{\pi(R_i), i = 1, 2, \ldots, n-m\} \quad \textit{and}$$
$$\nu(F) = \text{Max}\{\nu(R_i), i = 1, 2, \ldots, n-m\}.$$

Also,

$$\text{Ker}\, F \supset \text{Ker}\, R_1 + \text{Ker}\, R_2 + \cdots + \text{Ker}\, R_{n-m}.$$

Proof. We use Theorems 2.8 and 2.10 (whichever is appropriate in the one-step completions) as we build a simple diagonal completion R', which will satisfy the hypotheses of Theorem 1.8. □

4. Hermitian completions of band matrices with invertible maximal submatrices. In this section, we will establish Theorem 1.7, which says that a band matrix, with all its maximal submatrices being invertible, may be completed to a Hermitian matrix with any inertia that is consistent with Poincaré's inequalities.

The proof of Theorem 1.7 will begin by using the standard method for constructing Hermitian completions of band matrices using Theorem 2.10 at each one-step completion until the desired positivity and negativity are attained. Then we will quote Corollary 3.3.

Remark. In part (b) of Theorem 2.10, choosing z_1 or z_2 to complete $H(z)$ results in an increase in the positivity or negativity by only one and it is our choice as to which one increases. In part (a), the positivity (and negativity) of $H(z)$ is the same as the larger of the positivities (and negativities, respectively) of H_1 and H_3. This will enable us to "fine tune" the inertias as we do the sequence of one-step completions in the proofs of Theorems 1.7 and 1.11.

Proof of Theorem 1.7. Let R be a Hermitian m-band $n \times n$ matrix with *invertible maximal* submatrices $R_1, R_2, \ldots, R_{n-m}$ and positivity and negativity:

$$\pi^*(R) = \text{Max}\{\pi(R_i), i = 1, 2, \ldots, n-m\} \quad \text{and}$$
$$\nu^*(R) = \text{Max}\{\nu(R_i), i = 1, 2, \ldots, n-m\}.$$

Let $(\pi, \nu, \delta), \pi + \nu + \delta = n$ and $\pi^*(R) \leq \pi$ and $\nu^*(R) \leq \nu$ be given.

In the case, $(\pi^*(R), \nu^*(R)) = (\pi, \nu)$, Lemma 3.4 establishes this theorem.

We may now assume that $(\pi^*(R), \nu^*(R)) \neq (\pi, \nu)$.

We may construct successive simple diagonal completions, R to R' to $R'' = (R')' \cdots$ to $R^{(k)}$ using Theorem 2.10 (with $\operatorname{Ker} H_1 = 0 = \operatorname{Ker} H_3$ therein) repeatedly, at each one-step completion. Theorem 2.10 permits us to do this while having *all* the newly created maximal submatrices being invertible.

Furthermore, since Theorem 2.10 permits us to choose whether we want the positivity or the negativity to increase, and, in addition, since the increase is by at most one at each step, we can easily arrange it so that there is a successive simple diagonal completion $R^{(k)}$ with two maximal submatrices $R_i^{(k)}$ and $R_j^{(k)}$, such that $\pi^*(R_i^{(k)}) = \pi$ and $\nu^*(R_j^{(k)}) = \nu$, and In $R_t^{(k)} \leq (\pi, \nu, 0)$, for each maximal submatrix $R_t^{(k)}$ of $R^{(k)}$. (Note: $R_i^{(k)}$ and $R_j^{(k)}$ may be the same submatrix.)

We may then complete this band matrix $R^{(k)}$ to the desired completion F using Lemma 3.4. □

5. Hermitian completions of G-partial matrices, where G is a chordal graph. In this section, we will establish Theorems 1.10 and 1.11, which are our results on Hermitian completions of G-partial Hermitian matrices, when G is a chordal graph and the main diagonal is given.

OBSERVATION 5.1. Let H_2 be a submatrix of H_1, which in turn is a submatrix of a matrix H_0. Using the notation for identified subspaces listed in §1, we observe that if

$$\operatorname{Ker} H_0 \supset \operatorname{Ker} H_2, \quad \text{then } \operatorname{Ker} H_1 \supset \operatorname{Ker} H_2.$$

Remark. The standard method for constructing Hermitian completions of a G-partial Hermitian matrix, when G is a chordal graph and the main diagonal is specified, was presented in [10] and repeated in [11]. According to Lemma 4 of [10], ,the chordal graph structure provides a sequence of chordal graphs $\{G = G_0, G_1, \ldots, G_t\}$ such that going from each G_i to G_{i+1} corresponds to a one-step completion, i.e., if $\{R = F_0, F_1, \ldots, F_t = F\}$ are the (to be specified) partial Hermitian matrices whose graphs will be $\{G = G_0, G_1, \ldots, G_t\}$, then the unique new vertex v_i of G_i that is not in G_{i-1} will correspond to the unique pair of new off-diagonal entries, $z = r_{jk}$ and $z^* = r_{kj}$, which are *specified* in F_i but not in F_{i-1}; z and z^* together complete a bordered matrix $H(z)$ such that $H(z)$ is the *unique maximal* principal submatrix of F_i that is *not* also a principal submatrix of F_{i-1}.

In order to implement this induction process, it is sufficient that the set of properties that are to be transferred from the principal submatrices to the completed matrix are propagated by the one-step completions. For Theorem 1.10, the desired propagation properties, which are to be transferred from the principal submatrices to the newly completed matrices, are listed in Theorem 2.8.

Proof of Theorem 1.10. Let R be a G-partial Hermitian $n \times n$ matrix where G is a chordal graph and the hypotheses of this theorem are satisfied. Lemma 4 of [10] provides a sequence of chordal graphs $\{G = G_0, G_1, \ldots, G_t\}$. Let $\{R = F_0, F_1, \ldots, F_t\}$ be the (to be specified) partial Hermitian matrices whose graphs will be $\{G = G_0, G_1, \ldots, G_t\}$. Then going from each F_{k-1} to F_i is a one-step completion. Let $H(z)$ be the *unique maximal* principal submatrix of F_{k-1} that is not also a principal submatrix of F_k.

Induction hypotheses. We assume that $\pi^*(F_{k-1}) = \pi^*(R)$ and $\nu^*(F_{k-1}) = \nu^*(R)$ for the positivity and negativity of the $(k-1)$st partial Hermitian matrix F_{k-1} with chordal graph G_{k-1}. Also, for each principal submatrix S_j of a specified maximal principal submatrix M_i of F_{k-1}, we suppose that

$$(5.1) \qquad \operatorname{Ker} M_i \supset \operatorname{Ker} S_j .$$

Let $H(z)$ be the the bordered matrix, which is the *unique maximal* principal submatrix of F_{k-1} that is not also a principal submatrix of F_k, and let H_1, H_2, and H_3 be the principal submatrices of $H(z)$ and F_{k-1}, which overlap in the manner described by the bordered matrix hypotheses. Hence

$$(5.2) \qquad \pi^*(F_{i-1}) \geq \operatorname{Max}\{\pi(H_1), \pi(H_3)\} \text{ and } \nu^*(F_{i-1}) \geq \operatorname{Max}\{\nu(H_1), \nu(H_3)\}.$$

We note that (5.1) and Observation 5.1 imply that

$$\operatorname{Ker} H_1 \supset \operatorname{Ker} H_2 \text{ and } \operatorname{Ker} H_3 \supset \operatorname{Ker} H_2 .$$

Hence, Lemma 2.1(b) and (c), applied to both inclusions, yield

$$\delta(H_2) \leq \delta(H_1) \text{ and } \delta(H_2) \leq \delta(H_3),$$

and therefore Theorem 2.8 is applicable. Theorem 2.8 says that there is a number z_0 such that

$$(5.3) \qquad \operatorname{Ker} H(z_0) \supset \operatorname{Ker}(H_1 \oplus 1) + \operatorname{Ker}(1 \oplus H_3).$$

Also, using (5.2), we note that

$$\pi^*(F_i) = \operatorname{Max}\{\pi^*(F_{i-1}), \pi(H(z_0))\} = \operatorname{Max}\{\pi^*(F_{i-1}), \pi(H_1), \pi(H_3)\} = \pi^*(F_{i-1}),$$
$$\nu^*(F_i) = \operatorname{Max}\{\nu^*(F_{i-1}), \nu(H(z_0))\} = \operatorname{Max}\{\nu^*(F_{i-1}), \nu(H_1), \nu(H_3)\} = \nu^*(F_{i-1}).$$

Since $H(z_0)$ is the unique maximal principal submatrix of F_i that is *not* also a maximal principal submatrix of F_{i-1}, condition (5.1) for all the specified maximal principal submatrices of F_{i-1}, together with (5.3), implies that (5.1) is also valid for all the specified maximal principal submatrices of F_i. This and (2.3) show that the induction hypotheses will be satisfied for F_i.

We note that the final matrix F_t will be a full matrix with $\nu(F_t) = \nu^*(F_t) = \nu^*(R)$ and $\pi(F_t) = \pi^*(F_t) = \pi^*(R)$. Therefore, we set $F = F_t$, and this theorem is established. □

Remark. In the proof of Theorem 1.11, we will again use the standard method for constructing Hermitian completions of a G-partial Hermitian matrix when G is a chordal graph and the main diagonal is specified. For Theorem 1.11, the desired propagation property is that the newly created specified matrices be invertible; this will be achieved by Theorem 2.10. But this time we will only use this method until the specified values π and ν for the positivity and negativity are achieved. Then we will use Theorem 1.10 to complete the proof.

Proof of Theorem 1.11. Let R be a G-partial Hermitian $n \times n$ matrix, where G is a chordal graph, and the main diagonal of R is specified. We assume that all the specified principal submatrices of R are invertible. Let (π, ν, δ) be any nonnegative integers, such that $\pi + \nu + \delta = n$ and $\pi \geq \pi^*(R)$ and $\nu \geq \nu^*(R)$.

Again, according to Lemma 4 of [10], the chordal graph structure provides a sequence of chordal graphs $\{G = G_0, G_1, \dots, G_t\}$, such that going from each G_i to G_{i+1} corresponds to a one-step completion; i.e., if $\{R = F_0, F_1, \dots, F_t = F\}$ are the (to be specified) partial Hermitian matrices whose graphs will be $\{G = G_0, G_1, \dots, G_t\}$, then the unique new vertex v_i of G_i that is not in G_{i-1} will correspond to the unique pair of new off-diagonal entries $z = r_{jk}$ and $z^* = r_{kj}$, which are specified in F_i but not in $F_{i-1}; z$ and z^* together complete a bordered matrix $H(z)$, such that $H(z)$ is the *unique maximal* principal submatrix of F_i that is not also a principal submatrix of F_{i-1}.

Using Theorem 2.10 repeatedly at each one-step completion (with $\operatorname{Ker} H_1 = \underline{0} = \operatorname{Ker} H_3$ therein), we may construct the partial Hermitian completions $\{R = F_0, F_1, \dots, F_t\}$ of R, such that each principal submatrix of each F_i is invertible.

Furthermore, since Theorem 2.10 permits us to choose whether the positivity or the negativity increases, and, in addition, the increase is by *at most one* at each one-step completion, we can easily arrange it so that there will be a *first* partial Hermitian completion F_k with two principal submatrices A and B, such that $\pi(A) = \pi$ and $\nu(B) = \nu$ and $\operatorname{In} S \leq (\pi, \nu, 0)$, for each maximal principal submatrix S of F_k. (Note: A and B may be the same submatrix.)

Therefore, since F_k has the chordal graph G_k, and its main diagonal was specified, F_k may be completed to the desired Hermitian completion F with $\operatorname{In} F = (\pi, \nu, \delta)$ by Theorem 1.10. □

REFERENCES

[1] N. Cohen, C. R. Johnson, L. Rodman, and H. J. Woerdeman, *Ranks of completions of partial matrices*, Oper. Theory Adv. Appl., 40 (1989), pp. 165–185.

[2] J. Dancis, *The possible inertias for an Hermitian matrix and its principal submatrices*, Linear Algebra Appl., 85 (1987), pp. 121–151.

[3] ———, *On the inertias of symmetric matrices and bounded self-adjoint operators*, Linear Algebra Appl., 105 (1988), pp. 67–75.

[4] ———, *Bordered matrices*, Linear Algebra Appl., 128 (1990), pp. 117–132.

[5] ———, *Invertible completions of band matrices*, Linear Algebra Appl., 150 (1991), pp. 125–138.

[6] ———, *Positive semidefinite completions of partial Hermitian matrices*, Linear Algebra Appl., 175 (1992), pp. 97–114.

[7] H. Dym and I. Gohberg, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36 (1981), pp. 1–24.

[8] R. L. Ellis, I. Gohberg, and D. C. Lay, *Invertible selfadjoint extensions of band matrices and their entropy*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 483–500.

[9] R. L. Ellis and D. C. Lay, *Rank preserving extensions of band matrices*, Linear and Multilinear Algebra, 26 (1990), pp. 147–179.

[10] B. Grone, C. R. Johnson, E. M. Sa, and H. Wolkowicz, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.

[11] C. R. Johnson and L. Rodman, *Inertial possibilities for completions of partial Hermitean matrices*, Linear and Multilinear Algebra, 16 (1984), pp. 179–195.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 830–852, July 1993

016

LINEAR-ALGEBRAIC RESULTS ASSOCIATED WITH ANTIFERROMAGNETIC HEISENBERG CHAINS*

MAN-DUEN CHOI[†], JEFFREY S. ROSENTHAL[†][‡], AND PETER ROSENTHAL[†]

Abstract. This paper analyzes the $3^N \times 3^N$ selfadjoint matrix

$$H_N = \sum_{j=1}^{N-1} S_x^{(j)} S_x^{(j+1)} + S_y^{(j)} S_y^{(j+1)} + S_z^{(j)} S_z^{(j+1)},$$

the Hamiltonian of a spin-1, one-dimensional, Heisenberg antiferromagnet. Various results are obtained, including alternative representations of H_N, families of operators commuting with H_N, a complete description of the spin-0 subspace (which includes all one- and two-dimensional eigenspaces of H_N), and a proof that H_N must have an eigenvalue smaller than $-\frac{4}{3}(N-1)$.

Key words. antiferromagnets, Hamiltonian, eigenvalues, spin operators, spin chains

AMS subject classification. 81G10

1. Introduction.

There are several interesting, important, and difficult problems concerning the eigenvalues and eigenvectors of a certain sequence of Hermitian matrices that arise in the quantum-mechanical study of antiferromagnetic Heisenberg chains.

Let $\mathcal{V}$ be the three-dimensional inner-product space $\mathbb{C}^3$. The spin-1 operators (see, e.g., [8]) are the operators on $\mathcal{V}$ defined by

$$S_x = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad S_y = \frac{i}{\sqrt{2}}\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad S_z = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

with respect to an orthonormal basis, which we denote by $\{v_{-1}, v_0, v_1\}$. The choice of subscripts on the v_j is such that $S_z(v_j) = j\, v_j$ for $j = -1, 0, 1$. (Physicists write v_{-1} as $|-1\rangle$, v_0 as $|0\rangle$, and v_1 as $|1\rangle$.)

For each positive integer N, denote the N-fold tensor product of $\mathcal{V}$ with itself by $\mathcal{V}^{\otimes N}$. For each linear operator $F : \mathcal{V} \to \mathcal{V}$, let $F^{(j)} : \mathcal{V}^{\otimes N} \to \mathcal{V}^{\otimes N}$ be the operator

$$F^{(j)} = F_1 \otimes F_2 \otimes \cdots \otimes F_N$$

with $F_j = F$ and $F_k = I$ for $k \neq j$.

The operator we are concerned with is defined on $\mathcal{V}^{\otimes N}$ (for $N \geq 2$) by

$$H_N = \sum_{j=1}^{N-1} \left(S_x^{(j)} S_x^{(j+1)} + S_y^{(j)} S_y^{(j+1)} + S_z^{(j)} S_z^{(j+1)} \right).$$

*Received by the editors January 16, 1990, accepted for publication (in revised form) November 27, 1991. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

[†] Department of Mathematics, University of Toronto, Toronto, Canada M5S 1A1 (choi@math.toronto.edu and rosent@math.toronto.edu).

[‡] Present address, School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455 (rosentha@s5.math.umn.edu).

The operator H_N is the Hamiltonian of the spin-1, antiferromagnetic, Heisenberg (= isotropic), one-dimensional (= linear) spin chain with N sites. This Hamiltonian is of great importance in interpreting certain results in experimental physics (see [5] and [12]). Physically, such a Hamiltonian describes a crystal lattice of atoms of spin 1, in which all interactions take place along a preferred direction and are given by the dot-product of the spin vectors of all nearest neighbour pairs.

The operator H_N has been widely studied by solid state physicists (see [1], [3], [7], and references therein). There have also been numerical [4], [11] and experimental [5], [12] investigations of the eigenvalues and eigenvectors of H_N. Much of the recent work has been motivated by a conjecture of Haldane [9], [10]. Let λ_N^0 and λ_N^1 be the smallest and the second smallest eigenvalues of H_N. Haldane's conjecture may be stated as saying that $\lim_{N\to\infty} (\lambda_N^1 - \lambda_N^0) > 0$. While numerical and experimental work appears to support this statement, the conjecture remains unproven. Even the precise mathematical formulation of the conjecture is controversial; in [1] it is argued that the conjecture should be studied in an (inequivalent) "infinite chain" context.

We study H_N by direct, linear-algebraic methods in the present paper. We obtain a number of results that may provide insight into its underlying structure. Results presented here include alternative representations of H_N (Corollary 2.2 and Proposition 2.4), families of operators commuting with H_N (Theorems 3.3 and 3.5), properties of the spin-0 subspace (Propositions 4.3 and 4.5), a complete description of this subspace (Theorems 5.4 and 5.8), and a new proof that H_N must have an eigenvalue smaller than $-\frac{4}{3}(N-1)$ for all N (Theorem 6.7). The value $-\frac{4}{3}(N-1)$ is reasonably close to the estimate of $-1.40(N-1)$ for the smallest eigenvalue of H_N, which has been extrapolated from certain numerical approximations [11]. The spin-0 subspace is important in relation to Haldane's conjecture since it contains all one-dimensional (and two-dimensional) eigenspaces of H_N (see Corollary 4.2 and Proposition 6.8 below), and since it is shown in [1] that the eigenspace of H_N corresponding to λ_N^0 is one dimensional for even N. Our paper also includes simple, direct proofs of certain well-known facts about H_N, to make the presentation self-contained.

One way that physicists have varied the problem is by replacing H_N by the "periodic Hamiltonian,"

$$P_N = H_N + S_x^{(N)} S_x^{(1)} + S_y^{(N)} S_y^{(1)} + S_z^{(N)} S_z^{(1)}.$$

It is expected, on physical grounds, that the Haldane conjectures should be decided in the same way for P_N as for H_N. There are, however, no definitive results concerning P_N, either. We consider below both P_N and H_N. We often find it more natural (see §2) to study operators K_N and Q_N, which are equal to the negatives of H_N and P_N, respectively.

This paper is organized as follows. In §2, we derive several representations of H_N and P_N. In §3, we list a number of operators on $\mathcal{V}^{\otimes N}$ which commute with H_N and P_N. In §4, we use some of these operators to present the standard decomposition of H_N and P_N into direct sums, and in §5 we investigate in detail the spin-0 subspace of $\mathcal{V}^{\otimes N}$. Some bounds on the eigenvalues of H_N and P_N are obtained in §6. In §7, we describe the corresponding problems for spins other than 1 and state the more general Haldane conjecture, which remains one of most important open mathematical problems in solid state physics.

2. Other representations of the operators. The following theorem leads to new representations of H_N and P_N that we have found useful.

THEOREM 2.1. *There exists an orthonormal basis $\{e_1, e_2, e_3\}$ of $\mathcal{V}$ with respect to which the triple (S_x, S_y, S_z) has matrix representation (iR, iS, iT), where*

$$R = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad T = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Proof. Define the basis $\{e_1, e_2, e_3\}$ by

$$\begin{aligned} e_1 &= \frac{i}{\sqrt{2}}(v_{-1} - v_1), \\ e_2 &= \frac{-1}{\sqrt{2}}(v_{-1} + v_1), \\ e_3 &= i\, v_0. \end{aligned}$$

Then the stated matrix representations are easily checked. □

COROLLARY 2.2. *The operators H_N and P_N are equal, respectively, to the negatives of the operators K_N and Q_N defined (in the basis of $\mathcal{V}^{\otimes N}$ induced by $\{e_1, e_2, e_3\}$) by*

$$K_N = \sum_{j=1}^{N} R^{(j)}R^{(j+1)} + S^{(j)}S^{(j+1)} + T^{(j)}T^{(j+1)}$$

and

$$Q_N = K_N + R^{(N)}R^{(1)} + S^{(N)}S^{(1)} + T^{(N)}T^{(1)}.$$

Proof. This follows immediately from the preceding theorem. □

The representations given in Corollary 2.2 will be used throughout the rest of this paper. It should be observed that R, S, and T are skew-symmetric matrices, with $RS - SR = T$, $ST - TS = R$, $TR - RT = S$, and $R^2 + S^2 + T^2 = -2I$. Also, it is clear that the triple (R, S, T) is simultaneously unitarily equivalent to the triples (S, T, R) and (T, R, S). A possible interpretation of the matrices R, S, and T is suggested by the fact that for $v \in \mathbb{R}^3$, $Rv = e_1 \times v$, $Sv = e_2 \times v$, and $Tv = e_3 \times v$, where × indicates the ordinary cross-product (vector-product) on $\mathbb{R}^3$. We also have the following uniqueness property of the matrices R, S, and T.

PROPOSITION 2.3. *Let A, B, and C be three skew-Hermitian operators on $\mathbb{R}^3$, with $AB - BA = C$, $BC - CB = A$, and $CA - AC = B$. Then either $A = B = C = 0$, or there is an orthonormal basis $\{f_1, f_2, f_3\}$ of $\mathbb{R}^3$ such that in this basis A, B, and C have the matrix representations of R, S, and T above.*

Proof. Since $\det A = \det A^* = -\det A$, there is a vector $f_1 \in \mathbb{R}^3$ with $\|f_1\| = 1$ and $Af_1 = 0$. If $Cf_1 = 0$, then $Bf_1 = (CA - AC)f_1 = 0$, so with respect to any orthonormal basis containing f_1, each of A, B, and C are of the form

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -k \\ 0 & k & 0 \end{pmatrix}.$$

Hence, A, B, and C all commute, and thus are all zero. If $Cf_1 \neq 0$, then write $Cf_1 = cf_2$, with $\|f_2\| = 1$ and $c > 0$. Since

$$\langle Cf_1, f_1\rangle = \langle f_1, C^* f_1\rangle = \langle f_1, -Cf_1\rangle = -\langle Cf_1, f_1\rangle,$$

we have $f_1 \perp f_2$. Extend $\{f_1, f_2\}$ to an orthonormal basis $\{f_1, f_2, f_3\}$ of $\mathbb{R}^3$. With respect to this basis, we have

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -a \\ 0 & a & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & * & * \\ * & 0 & * \\ * & * & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & -c & 0 \\ c & 0 & * \\ 0 & * & 0 \end{pmatrix},$$

for some $a \in \mathbb{R}$. Direct computation then shows that

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -a \\ 0 & a & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & b \\ 0 & 0 & 0 \\ -b & 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & -c & 0 \\ c & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

for some $b \in \mathbb{R}$, with $ab = c$, $bc = a$, and $ca = b$. Replacing f_3 by $-f_3$, if necessary, we may assume $a > 0$. It then follows that $a = b = c = 1$, completing the proof. □

An interesting representation of the operator $K_2 = R \otimes R + S \otimes S + T \otimes T$ can be obtained by considering the space $M_3(\mathbb{C})$ of 3×3 complex matrices with inner product $\langle A, B \rangle = \text{trace } AB^*$. We identify $e_i \otimes e_j$ with the matrix unit E_{ij}, where $\{e_1, e_2, e_3\}$ is the orthonormal basis of Theorem 2.1; this induces a unitary equivalence of $\mathcal{V}^{\otimes 2}$ with $M_3(\mathbb{C})$. Under this equivalence, an operator on $\mathcal{V}^{\otimes 2}$ of the form $F \otimes G$ corresponds to the operator on $M_3(\mathbb{C})$ given by $A \mapsto FAG^t$, where F and G are written as matrices in the basis $\{e_1, e_2, e_3\}$.

PROPOSITION 2.4. *Under the above unitary equivalence of $\mathcal{V}^{\otimes 2}$ with $M_3(\mathbb{C})$, the operator K_2 corresponds to the operator Γ defined on $M_3(\mathbb{C})$ by*

$$\Gamma(A) = -A^t + (\operatorname{tr} A)I,$$

where A^t is the transpose of A, $\operatorname{tr} A$ is the trace of A, and I is the identity matrix.

Proof. We have

$$\Gamma(E_{ij}) = -E_{ji} + \delta_{ij} I = -E_{ji} + \delta_{ij}(E_{11} + E_{22} + E_{33}),$$

where δ is the Kronecker delta. On the other hand,

$$\begin{aligned} K_2(E_{ij}) &= RE_{ij}R^t + SE_{ij}S^t + TEijT^t \\ &= (E_{32} - E_{23})E_{ij}(E_{23} - E_{32}) + (E_{13} - E_{31})E_{ij}(E_{31} - E_{13}) \\ &\quad + (E_{21} - E_{12})E_{ij}(E_{12} - E_{21}), \end{aligned}$$

and the two expressions are easily seen to be equal. □

The spectrum of K_2 is very well known. It is not hard to compute this spectrum in any representation, but Proposition 2.4 makes it particularly easy.

COROLLARY 2.5. *The eigenvalues of K_2 are $2, 1$, and -1 with multiplicities $1, 3$, and 5, respectively. The corresponding eigenspaces of Γ are the subspaces, respectively, spanned by the identity matrix, the skew-symmetric matrices, and the symmetric matrices with trace 0.*

Proof. This follows immediately from the observation that $\Gamma(I) = 2I$, $\Gamma(A) = A$ if $A = -A^t$, and $\Gamma(A) = -A$ if $A = A^t$ and $\operatorname{tr} A = 0$. □

3. Operators commuting with the Hamiltonian. As mentioned in the introduction, it is shown in [1] that the eigenspaces of H_N and P_N corresponding to the smallest eigenvalue λ_N^0 have dimension 1 for N even. It is therefore important to study one-dimensional eigenspaces of H_N and P_N or, equivalently, of K_N and Q_N. Clearly, any such eigenspace is invariant under all operators commuting with K_N or Q_N. Hence, each operator A commuting with H_N or P_N puts a constraint on any eigenvector w of H_N or P_N corresponding to an eigenvalue of multiplicity 1, namely, that Aw must be a scalar multiple of w. This section describes two families of such operators. These families include operators which are known to physicists.

LEMMA 3.1. *Let F be a linear operator on $\mathcal{V}$. Then $F \otimes F$ commutes with K_2 if and only if, in the basis $\{e_1, e_2, e_3\}$ of Theorem 2.1, $FF^t = F^tF = \lambda I$, for some complex number λ, where I is the 3×3 identity matrix.*

Proof. We use Proposition 2.4. Under the equivalence given there, the operator $F \otimes F$ corresponds to multiplying an element of $M_3(\mathbb{C})$ on the left by F and on the right by F^t, where F is written as a matrix in the basis $\{e_1, e_2, e_3\}$. Hence $F \otimes F$ commutes with K_2 if and only if, for each $J \in M_3(\mathbb{C})$,

$$\Gamma(FJF^t) = F\Gamma(J)F^t,$$

i.e.,

$$-FJ^tF^t + \operatorname{tr}(FJF^t)\, I = F\,(-J^t + \operatorname{tr}(J)\, I)\,F^t,$$

which is true if and only if

$$(\dagger) \qquad \operatorname{tr}(FJF^t)\, I = \operatorname{tr}(J)\, FF^t.$$

If $FF^t = F^tF = \lambda I$, then $(\dagger)$ clearly holds. Conversely, if $(\dagger)$ holds, then setting $J = I$ shows that $FF^t = \frac{1}{3}\operatorname{tr}(FF^t)\, I$, a multiple of the identity. If $FF^t \neq 0$, we clearly have $F^tF = FF^t$. If $FF^t = 0$, then setting $J = (F^tF)^*$ in $(\dagger)$ shows that $\langle F^tF, F^tF\rangle = \operatorname{tr}\,(F^tFJ) = \operatorname{tr}\,(FJF^t) = 0$, so that $F^tF = 0$. □

Remark. The condition $FF^t = F^tF = \lambda I$ is not unitarily invariant, since in general $F^t \neq F^*$.

LEMMA 3.2. *Let F be a linear operator on $\mathcal{V}$. Let $\Theta_N^{(ij)} = R^{(i)}R^{(j)} + S^{(i)}S^{(j)} + T^{(i)}T^{(j)}$, where $i, j \in \{1, 2, \ldots, N\}$, $i \neq j$. Let $F^{\otimes N}$ denote the N-fold tensor product of F with itself. Then $F^{\otimes N}$ commutes with $\Theta_N^{(ij)}$ if and only if, in the basis $\{e_1, e_2, e_3\}$ of Theorem 2.1, $FF^t = F^tF = \lambda I$ for some complex number λ.*

Proof. Let Z be the operator on $\mathcal{V}^{\otimes N}$ defined by

$$Z(v_1 \otimes v_2 \otimes \cdots \otimes v_N) = v_{\sigma(1)} \otimes v_{\sigma(2)} \otimes \cdots \otimes v_{\sigma(N)},$$

extended by linearity, where σ is any fixed element of S_N (the symmetric group on N letters) with $\sigma(i) = 1$ and $\sigma(j) = 2$. Then Z clearly commutes with $F^{\otimes N}$, and $\Theta_N^{(ij)} = Z^{-1}\Theta_N^{(12)}Z$. Hence, $F^{\otimes N}$ commutes with $\Theta_N^{(ij)}$ if and only if it commutes with $\Theta_N^{(12)}$.

We can write $\Theta_N^{(12)} = K_2 \otimes I \otimes I \otimes \cdots \otimes I$. The lemma now follows easily from Lemma 3.1, by noting that $[F^{\otimes N}, \Theta_N^{(12)}] = [F \otimes F, K_2] \otimes F^{\otimes N-2}$. □

THEOREM 3.3. *Let $F^{\otimes N}$ be a linear operator on $\mathcal{V}$. Then $F^{\otimes N}$ commutes with K_N and with Q_N (and hence also with H_N and P_N) if and only if, in the basis $\{e_1, e_2, e_3\}$, $FF^t = F^tF = \lambda I$ for some complex number λ.*

Proof. Since each of K_N and Q_N are sums of $\Theta_N^{(ij)}$s, it follows immediately from Lemma 3.2 that K_N and Q_N commute with operators of the given form.

For the converse, let F be any operator on $\mathcal{V}$ such that $F^{\otimes N}$ commutes with K_N or Q_N. We shall show that $F^{\otimes N}$ must commute with $\Theta_N^{(1,2)}$; the result will then follow from Lemma 3.2.

Write the commutator of $F^{\otimes N}$ and $\Theta_N^{(12)}$ as

$$[F^{\otimes N}, \Theta_N^{(12)}] = \sum_j A_j \otimes B_j \otimes I^{\otimes N-2}$$

for some finite collection of operators A_j and B_j on $\mathcal{V}$. (In fact, it is easily seen that we only need three of each.) Now, recalling the definitions of K_N and Q_N, the only way we could possibly have

$$[F^{\otimes N}, K_N] = 0 \quad \text{or} \quad [F^{\otimes N}, Q_N] = 0$$

would be if for each j, either $A_j = I$ or $B_j = I$, and furthermore $\sum_j B_j = -\sum_j A_j$. This would imply that

$$[F^{\otimes N}, \Theta_N^{(12)}] = (I \otimes C - C \otimes I) \otimes I^{\otimes N-2}$$

for some operator C on $\mathcal{V}$. But by symmetry, we must also have

$$[F^{\otimes N}, \Theta_N^{(12)}] = (C \otimes I - I \otimes C) \otimes I^{\otimes N-2}.$$

It follows that $C = 0$, so that $F^{\otimes N}$ commutes with $\Theta_N^{(12)}$. Theorem 3.3 follows. □

LEMMA 3.4. *Let F be a linear operator on $\mathcal{V}$. Then $F \otimes I + I \otimes F$ commutes with K_2 if and only if, in the basis $\{e_1, e_2, e_3\}$, $F + F^t = \lambda I$ for some complex number λ.*

Proof. We again use Proposition 2.4. We have that

$$\begin{aligned}(F \otimes I + I \otimes F)\Gamma(A) &= -FA^t - A^tF^t + (\text{tr } A)F + (\text{tr } A)F^t \\ &= -FA^t - A^tF^t + (\text{tr } A)(F + F^t)\ ,\end{aligned}$$

and that

$$\begin{aligned}\Gamma\,(F \otimes I + I \otimes F)(A) &= -FA^t - A^tF^t + (\text{tr } FA)I + (\text{tr } AF^t)I \\ &= -FA^t - A^tF^t + (\text{tr } A(F + F^t)\,)I.\end{aligned}$$

The two operators commute if and only if the above two expressions are equal for every matrix A, and this is easily seen to be true if and only if $F + F^t$ is a multiple of the identity matrix. □

THEOREM 3.5. *Let F be a linear operator on $\mathcal{V}$. Then $\sum_{j=1}^N F^{(j)}$ commutes with K_N and with Q_N if and only if, in the basis $\{e_1, e_2, e_3\}$, $F + F^t = \lambda I$ for some complex number λ.*

Proof. This follows easily from Lemma 3.4, by techniques very similar to the proof of Lemma 3.2 and Theorem 3.3. □

Two other operators, both well known to physicists, deserve mention. Another operator in the commutant of K_N and of Q_N is the left-right symmetry L, of order 2, defined by

$$L(v_1 \otimes v_2 \otimes \cdots \otimes v_{N-1} \otimes v_N) = v_N \otimes v_{N-1} \otimes \cdots \otimes v_2 \otimes v_1,$$

extended by linearity. The commutant of Q_N also contains the rotation operator Π, of order N, defined by

$$\Pi(v_1 \otimes v_2 \otimes \cdots \otimes v_{N-1} \otimes v_N) = v_N \otimes v_1 \otimes v_2 \otimes \cdots \otimes v_{N-1},$$

extended by linearity. Since K_N and Q_N, and also Π and L, are real in the basis $\{e_1, e_2, e_3\}$, an element w of an eigenspace of K_N or Q_N of dimension 1 must satisfy $L(w) = \pm w$, and in the case of Q_N must also satisfy $\Pi(w) = \pm w$.

4. A decomposition of K_N and Q_N. Let k be an integer with $-N \leq k \leq N$. Let $\mathcal{M}_k^z$ be the eigenspace of $\sum_{j=1}^N S_z^{(j)}$ corresponding to the eigenvalue k. Similarly, let $\mathcal{M}_k^x$ and $\mathcal{M}_k^y$ be the eigenspaces of $\sum_{j=1}^N S_x^{(j)}$ and $\sum_{j=1}^N S_y^{(j)}$, respectively, corresponding to the eigenvalue k. Recall that in the basis $\{e_1, e_2, e_3\}$, the operators S_x, S_y, S_z have matrix representations R, S, T, respectively.

Using the orthonormal basis $\{v_{-1}, v_0, v_1\}$ of $\mathcal{V}$, note that

$$S_z^{(j)}(v_{r_1} \otimes \cdots \otimes v_{r_N}) = r_j(v_{r_1} \otimes \cdots \otimes v_{r_N}),$$

so that $\mathcal{M}_k^z$ is the span of

$$\left\{v_{r_1} \otimes v_{r_2} \otimes \cdots \otimes v_{r_N} \;\middle|\; r_1, \ldots, r_N \in \{-1, 0, 1\},\ r_1 + r_2 + \cdots + r_N = k\right\}.$$

This shows that $\mathcal{V}^{\otimes N} = \bigoplus_{k=-N}^N \mathcal{M}_k^z$, and by symmetry we have that $\mathcal{V}^{\otimes N} = \bigoplus_{k=-N}^N \mathcal{M}_k^x$ and $\mathcal{V}^{\otimes N} = \bigoplus_{k=-N}^N \mathcal{M}_k^y$.

The following proposition and its corollary are well known to physicists.

PROPOSITION 4.1. *Let q be one of x, y, or z. Then each $\mathcal{M}_k^q$ is invariant under K_N and under Q_N. Furthermore, every one-dimensional eigenspace of K_N or of Q_N is contained in $\mathcal{M}_0^q$.*

Proof. By symmetry, it suffices to consider the case $q = x$. The invariance of the $\mathcal{M}_k^x$ follows from the fact that $\sum_{j=1}^N R^{(j)}$ commutes with K_N and with Q_N, by Theorem 3.5. For the second statement, note that if $\mathcal{E}$ is a one-dimensional eigenspace, then since K_N and Q_N are real in the basis generated by $\{e_1, e_2, e_3\}$, we can choose $w \in \mathcal{E}$ with $w \neq 0$ and with w real in this basis. Suppose $w \in \mathcal{M}_k^x$. Then w is an eigenvector of $\sum_{j=1}^N R^{(j)}$ with eigenvalue $-ik$. But since $\sum_{j=1}^N R^{(j)}$ and w are real in the same basis, the eigenvalue must be real, so we must have $k = 0$. □

COROLLARY 4.2. *Any eigenspace of K_N or Q_N of dimension 1 is contained in the subspace $\mathcal{M}_0$ defined by*

$$\mathcal{M}_0 = \mathcal{M}_0^x \cap \mathcal{M}_0^y \cap \mathcal{M}_0^z.$$

Furthermore, $\mathcal{M}_0$ is invariant under K_N and Q_N.

Physically, $\mathcal{M}_0$ corresponds to the subspace of $\mathcal{V}^{\otimes N}$ with spin 0. If we had $\mathcal{M}_0 = \{0\}$, then K_N and Q_N would have no eigenspaces of multiplicity 1, contradicting [1]. However, it is known (see, e.g., [7]) that $\mathcal{M}_0$ is in fact fairly large. In the next section, we describe $\mathcal{M}_0$ precisely. We first present alternative characterizations of some of the subspaces considered in this section.

PROPOSITION 4.3. *The subspace $\mathcal{M}_0$ is equal to the intersection of the kernels of all operators on $\mathcal{V}^{\otimes N}$ of the form $\sum_{j=1}^N F^{(j)}$, where F is a linear operator on $\mathcal{V}$ with $F^t = -F$ in the basis $\{e_1, e_2, e_3\}$.*

Proof. Since each of the matrices R, S, and T of Theorem 2.1 are skew-symmetric, it is clear that $\mathcal{M}_0$ contains the given intersection. On the other hand, since every skew-symmetric matrix F is a linear combination of R, S, and T, it follows that $\sum_{j=1}^N F^{(j)}$ for such an F must annihilate every element in the kernels of each of $\sum_{j=1}^N R^{(j)}$, $\sum_{j=1}^N S^{(j)}$, and $\sum_{j=1}^N T^{(j)}$, and hence the given intersection must contain $\mathcal{M}_0$. □

PROPOSITION 4.4. *The subspaces $\mathcal{M}_0^x$, $\mathcal{M}_0^y$, and $\mathcal{M}_0^z$ are equal to the set of vectors in $\mathcal{V}^{\otimes N}$ left fixed by all operators of the form $F^{\otimes N}$, where F is a linear operator on $\mathcal{V}$ corresponding to a rotation about the e_1-axis, the e_2-axis, and the e_3-axis, respectively.*

Proof. We prove only the $\mathcal{M}_0^z$ part; the other statements then follow by permuting the roles of the e_is. It is easily checked that a rotation through an angle θ about the e_3-axis has eigenvectors v_{-1}, v_0, and v_1, with eigenvalues $e^{-i\theta}$, 1, and $e^{i\theta}$, respectively. Hence, any element of $\mathcal{M}_0^z$ is fixed by any N-fold tensor product of such a rotation.

Conversely, if a vector w is left invariant by all such N-fold products of rotations, choose any θ with θ/π irrational. Then the only way w can be fixed by the N-fold product of a rotation about the e_3-axis through that θ is if, in the basis generated by $\{v_{-1}, v_0, v_1\}$, each nonzero term in the expression for w has an equal number of v_{-1}s and v_1s. Hence $w \in \mathcal{M}_0^z$. □

PROPOSITION 4.5. *The subspace $\mathcal{M}_0$ is equal to the set of vectors in $\mathcal{V}^{\otimes N}$ which are left fixed by all operators of the form $F^{\otimes N}$, where F is a linear operator on $\mathcal{V}$ with $F^tF = I$ in the basis $\{e_1, e_2, e_3\}$, and with* $\det F = 1$.

Proof. If an element of $\mathcal{V}^{\otimes N}$ is left fixed by all such $F^{\otimes N}$, then by Proposition 4.5 it must be in each of $\mathcal{M}_0^x$, $\mathcal{M}_0^y$, and $\mathcal{M}_0^z$, and hence in $\mathcal{M}_0$.

For the converse, note that it follows by direct computation (by writing $F = A + iB$ and using the polar decomposition of A) that, in the basis $\{e_1, e_2, e_3\}$, we must have

$$F = O_1 \begin{pmatrix} \sqrt{1+b^2} & ib & 0 \\ -ib & \sqrt{1+b^2} & 0 \\ 0 & 0 & 1 \end{pmatrix} O_2$$

for some real number b and some real, orthogonal matrices O_1 and O_2. Furthermore, since $\det F = 1$, we can assume (by multiplying O_1 and O_2 by -1 if necessary) that $\det O_1 = \det O_2 = 1$. Now, every real, orthogonal matrix of determinant 1 can be written as a product of rotations about the e_1-axis and e_2-axis, so by Proposition 4.4 every element of $\mathcal{M}_0$ is fixed by every such matrix. As for the matrix

$$\begin{pmatrix} \sqrt{1+b^2} & ib & 0 \\ -ib & \sqrt{1+b^2} & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

it has eigenvectors v_{-1}, v_0, and v_1, with eigenvalues γ, 1, and γ^{-1}, respectively, where $\gamma = \sqrt{1+b^2} - b$, so it clearly fixes every element of $\mathcal{M}_0^z$. □

Remarks. (1) The proof above also shows that in fact $\mathcal{M}_0 = \mathcal{M}_0^x \cap \mathcal{M}_0^y$, etc. In other words, we can omit any one of the three sets being intersected in the definition of $\mathcal{M}_0$. However, we do not make use of this fact here.

(2) Part of Proposition 4.5 can be generalized to the statement that if $FF^t = F^tF = \lambda I$, then $F^{\otimes N}$ multiplies each element of $\mathcal{M}_0$ by $(\det F)^N$. For $\lambda \neq 0$, this follows immediately by considering $F \,/\, (\det F)$. For $\lambda = 0$, direct computation shows that up to real orthogonal matrices

$$F = \begin{pmatrix} 1 & i & 0 \\ -i & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

so that $F(v_{-1}) = 2v_{-1}$ and $F(v_0) = F(v_1) = 0$, from which it follows that $F^{\otimes N}$ annihilates each element of $\mathcal{M}_0^z$.

5. The structure of $\mathcal{M}_0$. In this section, we examine the subspaces $\mathcal{M}_0^x$, $\mathcal{M}_0^y$, $\mathcal{M}_0^z$, and $\mathcal{M}_0$ described above. Our main result (Theorems 5.4 and 5.8) is an explicit description of $\mathcal{M}_0$. Other, alternative descriptions have been obtained by the valence-bond basis approach; see [7].

We require some notation. For $\sigma \in S_N$ (the symmetric group on N letters), we define the linear operator Ω_σ on $\mathcal{V}^{\otimes N}$ by

$$\Omega_\sigma(u_1 \otimes \cdots \otimes u_N) = u_{\sigma(1)} \otimes \cdots \otimes u_{\sigma(N)},$$

extended by linearity.

Definition. Given a set $Y \subseteq \mathcal{V}^{\otimes N}$, the *permutation set* of Y is $P(Y) = \{\Omega_\sigma(y) \mid \sigma \in S_N, y \in Y\}$, and the *permutation-span* of Y, written P-sp(Y), is the linear span of $P(Y)$.

In this notation, we may write $\mathcal{M}_0^z$ as

$$\mathcal{M}_0^z = P\text{-sp}\Big\{(v_{-1} \otimes v_1)^{\otimes a} \otimes v_0^{\otimes b} \;\Big|\; 2a + b = N\Big\},$$

where, given a vector u, $u^{\otimes k}$ denotes the k-fold tensor product of u with itself.

In the basis $\{e_1, e_2, e_3\}$ of Theorem 2.1, this becomes

$$\mathcal{M}_0^z = P\text{-sp}\Big\{\big((-ie_1 - e_2) \otimes (ie_1 - e_2)\big)^{\otimes a} \otimes e_3^{\otimes b} \;\Big|\; 2a + b = N\Big\}.$$

Now note that this is the same as

$$P\text{-sp}\Big\{\big((ie_1 - e_2) \otimes (-ie_1 - e_2)\big) \otimes \big((-ie_1 - e_2) \otimes (ie_1 - e_2)\big)^{\otimes a-1} \otimes e_3^{\otimes b} \;\Big|\; 2a + b = N\Big\}.$$

By taking sums and differences of corresponding vectors in these two expressions, we obtain the "real" and "imaginary" parts of two tensor positions of these vectors (in the basis $\{e_{j_1} \otimes \cdots \otimes e_{j_N} \mid j_i \in \{1,2,3\}\}$), so we conclude that

$$\begin{aligned}\mathcal{M}_0^z = P\text{-sp}\Big(\Big\{&(e_1 \otimes e_1 + e_2 \otimes e_2) \otimes \big((-ie_1 - e_2) \otimes (ie_1 - e_2)\big)^{\otimes a-1} \\ &\qquad\qquad \otimes e_3^{\otimes b} \mid 2a + b = N\Big\} \\ &\cup \Big\{(e_1 \otimes e_2 - e_2 \otimes e_1) \otimes \big((-ie_1 - e_2) \otimes (ie_1 - e_2)\big)^{\otimes a-1} \\ &\qquad\qquad \otimes e_3^{\otimes b} \;\Big|\; 2a + b = N\Big\}\Big).\end{aligned}$$

Continuing in this way, we obtain, finally, that

$$(\dagger\dagger) \qquad \mathcal{M}_0^z = P\text{-sp}\Big\{\psi^{\otimes a} \otimes \chi^{\otimes b} \otimes e_3^{\otimes c} \;\Big|\; 2a + 2b + c = N\Big\},$$

where $\psi = e_1 \otimes e_2 - e_2 \otimes e_1$ and $\chi = e_1 \otimes e_1 + e_2 \otimes e_2$.

Now, note that $\psi \otimes \psi = \Omega_{\sigma_1}(\chi \otimes \chi) - \Omega_{\sigma_2}(\chi \otimes \chi)$, where σ_1 and σ_2 are the transpositions $\sigma_1 = (2\ 3)$ and $\sigma_2 = (2\ 4)$ (this can be checked simply by expanding both sides). This implies that we may assume in $(\dagger\dagger)$ that $a = 0$ or 1. Also, we can clearly replace χ by

$$\phi = \chi + e_3 \otimes e_3 = e_1 \otimes e_1 + e_2 \otimes e_2 + e_3 \otimes e_3,$$

so we have

$$(*)\ \mathcal{M}_0^z = P\text{-sp}\big(\{\phi^{\otimes a} \otimes e_3^{\otimes b} \mid 2a + b = N\} \cup \{\psi \otimes \phi^{\otimes a} \otimes e_3^{\otimes b} \mid 2 + 2a + b = N\}\big).$$

Let R, S, and T be as in Theorem 2.1. We can obtain $-S$ from T by interchanging the vectors e_2 and e_3 in the basis $\{e_1, e_2, e_3\}$. Similarly, we obtain $-R$ from T by interchanging e_1 and e_3. Since

$$\mathcal{M}_0^z = \ker\left(\sum_{j=1}^N T^{(j)}\right), \quad \mathcal{M}_0^x = \ker\left(\sum_{j=1}^N R^{(j)}\right), \quad \text{and} \quad \mathcal{M}_0^y = \ker\left(\sum_{j=1}^N S^{(j)}\right),$$

we have immediately from $(*)$ that

$$(**)\qquad \begin{aligned}\mathcal{M}_0^x &= P\text{-sp}\big(\{\phi^{\otimes a}\otimes e_1^{\otimes b} \mid 2a+b=N\}\\ &\qquad \cup\{(e_2\otimes e_3 - e_3\otimes e_2)\otimes\phi^{\otimes a}\otimes e_1^{\otimes b} \mid 2+2a+b=N\}\big)\end{aligned}$$

and

$$(***)\qquad \begin{aligned}\mathcal{M}_0^y &= P\text{-sp}\big(\{\phi^{\otimes a}\otimes e_2^{\otimes b} \mid 2a+b=N\}\\ &\qquad \cup\{(e_3\otimes e_1 - e_1\otimes e_3)\otimes\phi^{\otimes a}\otimes e_2^{\otimes b} \mid 2+2a+b=N\}\big).\end{aligned}$$

Since the vector ϕ will be important in what follows, we pause to note that ϕ will be seen to be the unique (up to scalar multiple) element of $\mathcal{M}_0$ for $N=2$, and that $K_2\phi = 2\phi$.

Having derived these expressions for $\mathcal{M}_0^x$, $\mathcal{M}_0^y$, and $\mathcal{M}_0^z$, we now consider their intersection, $\mathcal{M}_0$. We sometimes write $\mathcal{M}_{0,N}^x$ for $\mathcal{M}_0^x$, $\mathcal{M}_{0,N}^y$ for $\mathcal{M}_0^y$, $\mathcal{M}_{0,N}^z$ for $\mathcal{M}_0^z$, and $\mathcal{M}_{0,N}$ for $\mathcal{M}_0$, to emphasize the value of N under consideration.

LEMMA 5.1. (1) *Let q be one of x, y, and z. Let $u_1 \in \mathcal{M}_{0,N_1}^q$ and $u_2 \in \mathcal{M}_{0,N_2}^q$. Then $u_1\otimes u_2 \in \mathcal{M}_{0,N_1+N_2}^q$. The same result holds if we replace $\mathcal{M}_0^q$ by $\mathcal{M}_0$.*

(2) *Let $w \in \mathcal{M}_{0,N}$. Then P-sp$\{w\} \subseteq \mathcal{M}_{0,N}$.*

Proof. For Lemma 5.1(1), we have

$$\sum_{j=1}^{N_1+N_2} S_q^{(j)}(u_1\otimes u_2) = \sum_{j=1}^{N_1} S_q^{(j)}(u_1\otimes u_2) + \sum_{j=N_1+1}^{N_1+N_2} S_q^{(j)}(u_1\otimes u_2).$$

Since $u_1 \in \mathcal{M}_{0,N_1}^q$, the first of these two sums is zero. Since $u_2 \in \mathcal{M}_{0,N_2}^q$, the second is also zero. Hence

$$u_1\otimes u_2 \in \ker\left(\sum_{i=1}^{N_1+N_2} S_q^{(i)}\right) = \mathcal{M}_{0,N_1+N_2}^q.$$

The statement for $\mathcal{M}_0$ follows immediately. Lemma 5.1(2) is obvious. □

LEMMA 5.2. *Let $N \geq 2$ be even. Then*

$$\begin{aligned}\mathcal{M}_{0,N} =& P\text{-sp}\{\phi^{\otimes a}\otimes e_1^{\otimes b} \mid 2a+b=N\} \cap P\text{-sp}\{\phi^{\otimes a}\otimes e_2^{\otimes b} \mid 2a+b=N\}\\ &\cap P\text{-sp}\{\phi^{\otimes a}\otimes e_3^{\otimes b} \mid 2a+b=N\}.\end{aligned}$$

Proof. We use equations $(*)$, $(**)$, and $(***)$. Equation $(*)$ shows that

$$\mathcal{M}_{0,N}^z \subseteq \operatorname{span}\{e_{j_1}\otimes\cdots\otimes e_{j_N} \mid e_3 \text{ appears in an even number of positions}\}.$$

Combining this with equations $(**)$ and $(* * *)$ shows that

$$\mathcal{M}_{0,N} \subseteq \operatorname{span}\{e_{j_1} \otimes \cdots \otimes e_{j_N} \mid \text{each of } e_1, e_2, \text{ and } e_3 \text{ appears in an even number of positions}\}.$$

Now, an examination of equation $(*)$ shows that the only elements of $\mathcal{M}^z_{0,N}$ in this span are those in

$$P\text{-sp}\{\phi^{\otimes a} \otimes e_3^{\otimes b} \mid 2a+b=N\}.$$

Similarly, the only elements of $\mathcal{M}^x_{0,N}$ and $\mathcal{M}^y_{0,N}$ in this span are, respectively, those in

$$P\text{-sp}\{\phi^{\otimes a} \otimes e_1^{\otimes b} \mid 2a+b=N\}$$

and in

$$P\text{-sp}\{\phi^{\otimes a} \otimes e_2^{\otimes b} \mid 2a+b=N\}.$$

Since $\mathcal{M}_{0,N} = \mathcal{M}^x_{0,N} \cap \mathcal{M}^y_{0,N} \cap \mathcal{M}^z_{0,N}$, this completes the proof. □

COROLLARY 5.3. *Let $N \geq 2$ be even and let F be a linear operator on $\mathcal{V}$ which permutes the basis $\{e_1, e_2, e_3\}$. Then $F^{\otimes N}$ fixes $\mathcal{M}_{0,N}$.*

Proof. Since every element of $P\text{-sp}\{\phi^{\otimes a} \otimes e_3^{\otimes b} \mid 2a+b=N\}$ is unchanged upon interchanging e_1 and e_2, the corollary is true when F arises from the transposition (1 2). Similarly, the corollary is true when F arises from the transposition (2 3). Since (1 2) and (2 3) generate S_3, we are done. (This corollary also follows from Proposition 4.5.) □

THEOREM 5.4. *Let $N \geq 2$ be even, and let $\phi = e_1 \otimes e_1 + e_2 \otimes e_2 + e_3 \otimes e_3$. Then*

$$\mathcal{M}_{0,N} = P\text{-sp}\{\phi^{\otimes N/2}\}.$$

Proof. Equations $(*)$, $(**)$, and $(* * *)$ show $\mathcal{M}_{0,N} \supseteq P\text{-sp}\{\phi^{\otimes N/2}\}$. Conversely, given $w \in \mathcal{M}_{0,N}$, we proceed to show that $w \in P\text{-sp}\{\phi^{\otimes N/2}\}$.

By Lemma 5.2,

$$w \in P\text{-sp}\{\phi^{\otimes a} \otimes e_3^{\otimes b} \mid 2a+b=N\},$$

so we can write

$$w = \sum_{\substack{b=0 \\ b \text{ even}}}^{N/2} \sum_{\sigma \in S_N} \alpha_{b,\sigma}\, \Omega_\sigma \left(\phi^{\otimes (N-b)/2} \otimes e_3^{\otimes b}\right),$$

where the coefficients $\alpha_{b,\sigma}$ are complex numbers. Let $b_{\max}$ be the largest value of b for which some $\alpha_{b,\sigma} \neq 0$, and let us suppose our expression for w is such that $b_{\max}$ is as small as possible. We wish to show that this minimal $b_{\max}$ is 0, for then $w \in P\text{-sp}\{\phi^{\otimes N/2}\}$, as desired.

Suppose $b_{\max} > 0$, so $b_{\max} \geq 2$. By Corollary 5.3, w is invariant upon interchanging e_1 and e_3, so we have

$$(\diamond) \qquad w = \sum_{\substack{b=0 \\ b \text{ even}}}^{N} \sum_{\sigma \in S_N} \alpha_{b,\sigma}\, \Omega_\sigma \left(\phi^{\otimes (N-b)/2} \otimes e_1^{\otimes b}\right).$$

Define the operator Φ_N on $\mathcal{V}^{\otimes N}$ to be the orthogonal projection onto $\mathcal{M}^z_{0,N}$. Thinking in the basis $\{v_{-1}, v_0, v_1\}$ and using Lemma 5.1(1), it is clear that if $w_1 \in \mathcal{M}^z_{0,N_1}$, and if $w_2 \in \mathcal{V}^{\otimes N_2}$, then $\Phi_{N_1+N_2}(w_1 \otimes w_2) = w_1 \otimes \Phi_{N_2}(w_2)$. Now, $\Phi_N(w) = w$, so we have

$$\begin{aligned} w &= \Phi_N\left(\sum_{\substack{b=0\\ b \text{ even}}}^{N} \sum_{\sigma \in S_N} \alpha_{b,\sigma}\Omega_\sigma \left(\phi^{\otimes(N-b)/2} \otimes e_1^{\otimes b}\right)\right) \\ &= \sum_{\substack{b=0\\ b \text{ even}}}^{N} \sum_{\sigma \in S_N} \alpha_{b,\sigma}\Phi_N \left(\Omega_\sigma \left(\phi^{\otimes(N-b)/2} \otimes e_1^{\otimes b}\right)\right) \\ &= \sum_{\substack{b=0\\ b \text{ even}}}^{N} \sum_{\sigma \in S_N} \alpha_{b,\sigma}\Omega_\sigma \left(\Phi_N \left(\phi^{\otimes(N-b)/2} \otimes e_1^{\otimes b}\right)\right) \\ &= \sum_{\substack{b=0\\ b \text{ even}}}^{N} \sum_{\sigma \in S_N} \alpha_{b,\sigma}\Omega_\sigma \left(\phi^{\otimes(N-b)/2} \otimes \Phi_b\left(e_1^{\otimes b}\right)\right), \end{aligned}$$

the last equality following from the fact that $\phi^{\otimes(N-b)/2} \in \mathcal{M}^z_{0,N-b}$. Now,

$$\begin{aligned} \Phi_b\left(e_1^{\otimes b}\right) &= \Phi_b\left(\left(\frac{i}{\sqrt{2}}(v_{-1} - v_1)\right)^{\otimes b}\right) \\ &= \left(-\frac{1}{2}\right)^{b/2} \frac{1}{((b/2)!)^2} \sum_{\sigma \in S_b} \Omega_\sigma\left((-v_{-1} \otimes v_1)^{\otimes b/2}\right) \\ &= \left(\frac{1}{2}\right)^{b/2} \frac{1}{((b/2)!)^2} \sum_{\sigma \in S_b} \Omega_\sigma\left((v_{-1} \otimes v_1)^{\otimes b/2}\right) \\ &= \left(\frac{1}{2}\right)^{b/2} \frac{1}{((b/2)!)^2} \sum_{\sigma \in S_b} \left(\frac{1}{2}\right)^{b/2} \Omega_\sigma\left((v_{-1} \otimes v_1 + v_1 \otimes v_{-1})^{\otimes b/2}\right) \\ &= \frac{1}{2^b} \frac{1}{((b/2)!)^2} \sum_{\sigma \in S_b} \Omega_\sigma\left((\phi - e_3 \otimes e_3)^{\otimes b/2}\right) \\ &= \frac{(-1)^{b/2}\, b!}{2^b\, ((b/2)!)^2}\, e_3^{\otimes b} + \cdots, \end{aligned}$$

where the "$\cdots$" indicates terms involving at least one ϕ, and hence no more than $b-2$ e_3's. Hence

$$w = \sum_{\substack{b=0\\ b \text{ even}}}^{N} \sum_{\sigma \in S_N} \alpha_{b,\sigma}\, \Omega_\sigma \left(\phi^{\otimes(N-b)/2} \otimes \left(\beta_b\, e_3^{\otimes b} + \cdots\right)\right), \quad \text{where } \beta_b = \frac{(-1)^{b/2}\, b!}{2^b\, ((b/2)!)^2}.$$

Note that for any $b \geq 2$, β_b is not 1. Subtracting this expression for w from $\beta_{b_{\max}}$ times ($\diamond$) yields

$$\left(\beta_{b_{\max}} - 1\right) w = \sum_{\substack{b=0\\ b \text{ even}}}^{N} \sum_{\sigma \in S_N} \alpha_{b,\sigma}\, \Omega_\sigma \left(\phi^{\otimes(N-b)/2} \otimes \left(\left(\beta_{b_{\max}} - \beta_b\right) e_3^{\otimes b} + \cdots\right)\right).$$

Dividing this last expression by $(\beta_{b_{\max}} - 1)$ yields an expression for w which has terms corresponding only to values of b strictly less than $b_{\max}$. This contradicts the assumption that our $b_{\max}$ in $(\diamond)$ was minimal. Hence, it must have been that the true minimal $b_{\max}$ was 0, and therefore that $w \in P\text{-sp}\{\phi^{\otimes N/2}\}$. □

Remark. Note that the vector $\Omega_\sigma\left(\phi^{\otimes N/2}\right)$ depends on σ only through the unordered, indistinguishable pairs $\{\sigma(1),\sigma(2)\},\{\sigma(3),\sigma(4)\},\ldots,\{\sigma(N-1),\sigma(N)\}$. Hence, when considering a vector of the form $\Omega_\sigma\left(\phi^{\otimes N/2}\right)$, we can assume without loss of generality that σ is *canonical* in the sense that $\sigma(2j-1) < \sigma(2j)$ for $1 \le j \le N/2$, and that $\sigma(2) < \sigma(4) < \cdots < \sigma(N)$.

To determine the structure of $\mathcal{M}_{0,N}$ for N odd, we require two additional linear operators. For $N \ge 1$, for any $w \in \mathcal{V}^{\otimes N-1}$, and (i,j,k) any cyclic permutation of $(1,2,3)$, we define $\Psi : \mathcal{V}^{\otimes N} \to \mathcal{V}^{\otimes N+1}$ by

$$\Psi(w \otimes e_i) = \tfrac{1}{2}\left(w \otimes (e_j \otimes e_k - e_k \otimes e_j)\right),$$

and define $\Lambda : \mathcal{V}^{\otimes N+1} \to \mathcal{V}^{\otimes N}$ by

$$\begin{aligned}\Lambda(w \otimes e_i \otimes e_i) &= 0,\\ \Lambda(w \otimes e_i \otimes e_j) &= w \otimes e_k, \quad \text{and}\\ \Lambda(w \otimes e_j \otimes e_i) &= -w \otimes e_k.\end{aligned}$$

We extend Ψ and Λ by linearity.

LEMMA 5.5. (1) $\Lambda \circ \Psi$ *is the identity on* $\mathcal{V}^{\otimes N}$.

(2) $\Psi\left(\mathcal{M}_{0,N}\right) \subseteq \mathcal{M}_{0,N+1}$.

(3) $\Lambda\left(\mathcal{M}_{0,N+1}\right) \subseteq \mathcal{M}_{0,N}$.

Proof. Lemma 5.5(1) is obvious. For Lemma 5.5(2), note that it suffices to show that

$$\left(\sum_{j=1}^{N+1} X^{(j)}\right)\Psi = \Psi\left(\sum_{j=1}^{N} X^{(j)}\right),$$

where $X = R$, S, and T, as defined in Theorem 2.1. By symmetry, it suffices to consider the case $X = R$, and clearly we need only show

$$\left(R^{(N)} + R^{(N+1)}\right)\Psi = \Psi\, R^{(N)}.$$

This follows by direct computation on elements of the form $w \otimes e_i$, where $i = 1,2,3$. Similarly, for Lemma 5.5(3), it suffices to show

$$R^{(N)}\Lambda = \Lambda\left(R^{(N)} + R^{(N+1)}\right),$$

and this follows by direct computation on elements of the form $w \otimes e_i \otimes e_j$, where $i,j = 1,2,3$. □

PROPOSITION 5.6. *For any* $N \ge 2$,

$$\mathcal{M}_{0,N} = \Lambda\left(\mathcal{M}_{0,N+1}\right).$$

Proof. By Lemma 5.5(3), it suffices to show $\Lambda\left(\mathcal{M}_{0,N+1}\right) \supseteq \mathcal{M}_{0,N}$. Using (2) and (1) of Lemma 5.5, we have

$$\Lambda\left(\mathcal{M}_{0,N+1}\right) \supseteq \Lambda\left(\Psi\left(\mathcal{M}_{0,N}\right)\right) = \mathcal{M}_{0,N},$$

completing the proof. □

LEMMA 5.7. *Let* $\Delta = \sum_{\sigma\in S_3}(\operatorname{sgn}\sigma)\, e_{\sigma(1)}\otimes e_{\sigma(2)}\otimes e_{\sigma(3)}$ *, where* S_3 *is the symmetric group on three letters, and* sgn σ *is the sign of the permutation* σ*. Then* $\Delta \in \mathcal{M}_{0,3}$*.*

Proof. We first show that $\left(\sum_{j=1}^3 R^{(j)}\right)(\Delta) = 0$. Note that

$$\begin{aligned}\left(R^{(1)} + R^{(2)} + R^{(3)}\right)(e_1\otimes e_2\otimes e_3) &= 0 \;+\; e_1\otimes e_3\otimes e_3 \;-\; e_1\otimes e_2\otimes e_2 \\ &= \left(R^{(1)} + R^{(3)} + R^{(2)}\right)(e_1\otimes e_3\otimes e_2).\end{aligned}$$

Hence $\left(\sum_{j=1}^3 R^{(j)}\right)(e_1\otimes e_2\otimes e_3 - e_1\otimes e_3\otimes e_2) = 0$. The other terms in $\left(\sum_{j=1}^3 R^{(j)}\right)(\Delta)$ cancel similarly. Hence $\Delta \in \mathcal{M}_0^x$. Similarly, $\Delta \in \mathcal{M}_0^y$ and $\Delta \in \mathcal{M}_0^z$. □

THEOREM 5.8. *Let* $N \geq 3$ *be odd. Then*

$$\mathcal{M}_{0,N} = P\text{-sp}\left\{\phi^{\otimes(N-3)/2}\otimes\Delta\right\},$$

with $\Delta = \sum_{\sigma\in S_3}(\operatorname{sgn}\sigma)\, e_{\sigma(1)}\otimes e_{\sigma(2)}\otimes e_{\sigma(3)}$.

Proof. Theorem 5.4 and Lemmas 5.1 and 5.7 show

$$\mathcal{M}_{0,N} \supseteq P\text{-sp}\left\{\phi^{\otimes(N-3)/2}\otimes\Delta\right\}.$$

Conversely, Proposition 5.6 and Theorem 5.4 show that

$$\begin{aligned}\mathcal{M}_{0,N} &= \Lambda\left(\mathcal{M}_{0,N+1}\right) \\ &= \Lambda\left(P\text{-sp}\left\{\phi^{\otimes(N+1)/2}\right\}\right),\end{aligned}$$

so it suffices to show

$$\Lambda\left(\Omega_\sigma\left(\phi^{\otimes(N+1)/2}\right)\right) \in P\text{-sp}\left\{\phi^{\otimes(N-3)/2}\otimes\Delta\right\},$$

for any $\sigma \in S_{N+1}$. By the remark following Theorem 5.4, we can assume that σ is canonical, and, in particular, that $\sigma(N+1) = N+1$, and either $\sigma(N) = N$ or $\sigma(N-1) = N$. In the first case, $\Lambda\left(\Omega_\sigma\left(\phi^{\otimes(N+1)/2}\right)\right)$ is zero, while in the second case

$$\Lambda\left(\Omega_\sigma\left(\phi^{\otimes(N+1)/2}\right)\right) = \Omega_\tau\left\{\phi^{\otimes(N-3)/2}\otimes\Delta\right\},$$

where τ equals σ restricted to $\{1,2,\ldots,N\}$. This completes the proof. □

Remark. The proof above actually shows that

$$P\text{-sp}\left\{\phi^{\otimes(N-3)/2}\otimes\Delta\right\} = \operatorname{span}\left\{\Omega_\sigma\left(\phi^{\otimes(N-3)/2}\otimes\Delta\right) \mid \sigma\in S_N,\ \sigma(N-1) = N\right\}.$$

In other words, we need only consider those σ with $\sigma(N-1) = N$. Since Δ is skew-symmetric upon interchanging its last two tensor positions, we may instead assume $\sigma(N) = N$. Combining this with reasoning as in the remark following Theorem 5.4, we see that we can assume σ is canonical in the sense that $\sigma(N) = N$, $\sigma(2j-1) < \sigma(2j)$ for $1 \leq j \leq (N-1)/2$, and $\sigma(2) < \sigma(4) < \cdots \sigma(N-3)$. (Note that we cannot assume $\sigma(N-3) < \sigma(N-1)$, since the unordered pair $\{\sigma(N-2), \sigma(N-1)\}$ is special and must be allowed to occur anywhere.)

We now turn our attention to the dimension of $\mathcal{M}_{0,N}$. By Proposition 5.6, this dimension is an increasing function of N. It is shown in [6] (see [7]) that for N even,

$$\dim\mathcal{M}_{0,N} = \sum_{m=0}^{N/2}\frac{N!}{(m!)^2\,(N-2m)!} - \sum_{m=0}^{(N/2)-1}\frac{N!}{m!(m+1)!(N-2m-1)!}.$$

This expression, while exact, is difficult to work with. Furthermore, it holds for even N only.

We present here some upper bounds on $\dim \mathcal{M}_{0,N}$ in closed form, which follow directly from the results of this section.

PROPOSITION 5.9. *Let $N \geq 4$ be even. Then*

$$\dim \mathcal{M}_{0,N} \leq (N-1) \dim \mathcal{M}_{0,N-2}.$$

Proof. Recall that

$$\mathcal{M}_{0,N} = \operatorname{span}\left\{\Omega_\sigma\left(\phi^{\otimes N/2}\right)\right\}.$$

Furthermore, by the remark following Theorem 5.4, we need only consider canonical σ, so that $\sigma(N) = N$. There are then $(N-1)$ possible values of $\sigma(N-1)$. If we "delete" tensor positions N and $\sigma(N-1)$, we are left with an element of $\mathcal{M}_{0,N-2}$, so that

$$\dim \operatorname{span}\left\{\Omega_\sigma\left(\phi^{\otimes N/2}\right) \,\middle|\, \sigma(N) = N, \sigma(N-1) = j\right\} = \dim \mathcal{M}_{0,N-2}$$

for $j = 1, 2, \ldots, N-1$. The result follows. □

COROLLARY 5.10. *Let $N \geq 2$ be even. Then*

$$\dim \mathcal{M}_{0,N} \leq (N-1)(N-3)\cdots \times 3 \cdot 1 = \frac{N!}{(\frac{N}{2})!\,2^N}.$$

Furthermore, equality holds for $N = 2$, 4, or 6.

Proof. For $N = 2$, clearly $\dim \mathcal{M}_{0,2} = \dim\{\phi\} = 1$. The inequality now follows from Proposition 5.9 by induction. For $N = 6$, to show equality it suffices to show that the set

$$\left\{\Omega_\sigma\left(\phi^{\otimes 3}\right) \,\middle|\, \sigma \in S_6,\ \sigma \text{ canonical}\right\}$$

is linearly independent. This follows from the observation that for σ_1 and σ_2 canonical,

$$\left\langle \Omega_{\sigma_1}\left(\phi^{\otimes 3}\right),\ \Omega_{\sigma_2}\left(e_1 \otimes e_1 \otimes e_2 \otimes e_2 \otimes e_3 \otimes e_3\right)\right\rangle$$

is 1 or 0 as $\sigma_1 = \sigma_2$ or $\sigma_1 \neq \sigma_2$. The proof of equality for $N = 4$ is similar, involving $e_1 \otimes e_1 \otimes e_2 \otimes e_2$ in place of $e_1 \otimes e_1 \otimes e_2 \otimes e_2 \otimes e_3 \otimes e_3$. □

PROPOSITION 5.11. *Let $N \geq 3$ be odd. Then*

$$\dim \mathcal{M}_{0,N} \leq \left(\frac{N-1}{2}\right)(N-2)(N-4)\cdots \times 3 \cdot 1.$$

Furthermore, equality holds for $N = 3$ and $N = 5$.

Proof. The inequality follows immediately from the fact that the number of canonical elements of S_N, in the sense of the remark following Theorem 5.8, is precisely $((N-1)/2)\,(N-2)(N-4)\cdots \times 3 \cdot 1$. For $N = 3$, clearly $\dim \mathcal{M}_{0,3} = \dim\{\Delta\} = 1$. For equality when $N = 5$, note that if σ_1 and σ_2 are two canonical elements of S_5, then

$$\left\langle \Omega_{\sigma_1}\left(\phi \otimes \Delta\right),\ \Omega_{\sigma_2}\left(e_1 \otimes e_1 \otimes e_2 \otimes e_3 \otimes e_1\right)\right\rangle$$

is 1 or 0 as $\sigma_1 = \sigma_2$ or $\sigma_1 \neq \sigma_2$. Hence the set

$$\left\{\Omega_\sigma\left(\phi \otimes \Delta\right) \,\middle|\, \sigma \in S_5,\ \sigma \text{ canonical}\right\}$$

is linearly independent. □

6. Some special vectors: bounds on eigenvalues of K_N and Q_N. As mentioned in the introduction, Haldane's conjecture [9], [10] involves the lowest eigenvalues of H_N and P_N, or equivalently the highest eigenvalues of $K_N = -H_N$ and $Q_N = -P_N$. Numerical work by physicists (see, e.g., [11]) suggests that for large N the lowest eigenvalues of P_N are approximately $-1.4015\,N$, with a difference (or "gap") of about 0.41 between the two lowest ones, and it is believed that the corresponding values for H_N are similar. In this section, we present a few bounds on eigenvalues of K_N and Q_N, and show (Theorem 6.7) that H_N and P_N have eigenvalues lower than $-\frac{4}{3}(N-1)$ and about $-\frac{4}{3}N$, respectively.

A similar result is obtained by a valence-bond approach in [2]. There, vectors $\Omega_{\alpha\beta}(\alpha,\beta \in \{1,2\})$ are constructed, and the expected values of H_N are considered. (The vectors $\Omega_{\alpha\beta}$ are defined more explicitly in [3].) Our approach will be somewhat similar, in that we will construct vectors $\omega_{j,N}$ and consider expressions like $\langle H_N(\omega_{j,N}), \omega_{j,N}\rangle$. Our vectors $\omega_{j,N}$ are different than the vectors $\Omega_{\alpha,\beta}$, and are defined in a quite different way. However, there are certain strong connections. For example, the identity $\Omega_{12} + (-1)^N\Omega_{21} = 2\omega_{0,N}$ appears to hold in general.

We begin with an obvious bound.

PROPOSITION 6.1. *The spectrum of K_N is contained in $[-(N-1), 2(N-1)]$, and that of Q_N is contained in $[-N, 2N]$.*

Proof. Recall from Corollary 2.5 that the spectrum of K_2 is $\{-1,1,2\}$. This is clearly the same as the spectrum of $\Theta_N^{(ij)}$ as defined in Lemma 3.2. Proposition 6.1 follows from the fact that K_N is the sum of $N-1$ operators of the form $\Theta_N^{(ij)}$, and Q_N is the sum of N operators of the form $\Theta_N^{(ij)}$. □

The smallest eigenvalues above are not relevant to Haldane's conjecture, but they are important in the study of ferromagnets, where the Hamiltonian is the negative of the operator H_N presented here. To examine these, we require the following definitions, also to be used elsewhere in this section.

DEFINITIONS. A vector $u \in \mathcal{V}^{\otimes N}$ is

(a) symmetric,
(b) skew-symmetric,
(c) scalar, or
(d) of trace 0,

in tensor positions 1 and 2, if

(a) $\Omega_{(12)}(u) = u$,
(b) $\Omega_{(12)}(u) = -u$,
(c) $u = \phi \otimes u_0$ for some $u_0 \in \mathcal{V}^{\otimes N-2}$, where $\phi = e_1 \otimes e_1 + e_2 \otimes e_2 + e_3 \otimes e_3$, or
(d) $u = \sum_{ij} e_i \otimes e_j \otimes u_{ij}$, with $u_{11} + u_{22} + u_{33} = 0$.

(Here Ω permutes the tensor positions, and is defined in the beginning of §5.) The vector u has one of the above properties in tensor positions k and l $(k < l)$ if $\Omega_{(2l)(1k)}(u)$ has the corresponding property in tensor positions 1 and 2. We say u is *scalar plus skew-symmetric* in tensor positions k and l if $u + \Omega_{(kl)}(u)$ is scalar in tensor positions k and l, or equivalently if $u = u_1 + u_2$ with u_1 scalar and u_2 skew-symmetric in tensor positions k and l.

By Corollary 2.5, u is scalar in tensor positions k and l if and only if $\Theta_N^{(kl)}(u) = 2u$, u is skew-symmetric in tensor positions k and l if and only if $\Theta_N^{(kl)}(u) = u$, and u is symmetric and of trace 0 in tensor positions k and l if and only if $\Theta_N^{(kl)}(u) = -u$.

The following proposition, important in the study of ferromagnets, is well known to physicists.

PROPOSITION 6.2. *The eigenspace of K_N corresponding to the eigenvalue $-(N-1)$ is the same as the eigenspace of Q_N corresponding to the eigenvalue $-N$. This common eigenspace has dimension $2N+1$.*

Proof. For a vector $w \in \mathcal{V}^{\otimes N}$ to be in the eigenspace for K_N, it must be that $\Theta_N^{(i,i+1)}(w) = -w$ for all adjacent positions i and $i+1$. From the above discussion, this happens precisely when w is symmetric and of trace 0 in each two adjacent positions. This condition translates into the following. If we write

$$w = \sum_h \alpha_h \, e_{h(1)} \otimes e_{h(2)} \otimes \cdots \otimes e_{h(N)},$$

where the sum is taken over all functions $h : \{1,2,\ldots,N\} \to \{1,2,3\}$ and the coefficients α_h are complex numbers, then we must have:

(A) $\alpha_{h_1} = \alpha_{h_2}$ whenever $h_1 = h_2 \circ \tau$ for some $\tau = (i\ i{+}1) \in S_N$, and

(B) $\alpha_{h_1} + \alpha_{h_2} + \alpha_{h_3} = 0$ whenever for some i, $h_1(i) = h_1(i+1) = 1$, $h_2(i) = h_2(i+1) = 2$, $h_3(i) = h_3(i+1) = 3$, and $h_1(k) = h_2(k) = h_3(k)$ for $k \neq i, i+1$.

In the case of Q_N, we require the same conditions, but must also allow the pair $(N,1)$ in place of $(i,i+1)$. In either case, conditions (A) and (B) are easily seen to be equivalent to

(A$'$) $\alpha_{h_1} = \alpha_{h_2}$ whenever $h_1 = h_2 \circ \tau$ for any $\tau \in S_N$, and

(B$'$) $\alpha_{h_1} + \alpha_{h_2} + \alpha_{h_3} = 0$ whenever, for some $i \neq j$, $h_1(i) = h_1(j) = 1$, $h_2(i) = h_2(j) = 2$, $h_3(i) = h_3(j) = 3$, and $h_1(k) = h_2(k) = h_3(k)$ for $k \neq i, j$.

Conditions (A$'$) and (B$'$) make it clear that we can choose arbitrarily the coefficients α_h for "primitive" functions h, i.e., functions h with at most one element in $h^{-1}\{3\}$, and with $h(1) \leq h(2) \leq \cdots \leq h(N)$. The coefficients α_h with at most one element in $h^{-1}\{3\}$, but with arbitrary ordering, are then forced by condition (A$'$). The coefficients α_h with h having more and more elements in $h^{-1}\{3\}$ are then forced in turn by condition (B$'$). Under such a procedure, condition (A$'$) is satisfied automatically. Since there are $2N+1$ "primitive" functions h, this is the dimension in question. □

As a simple example of vectors in the eigenspace described above, note that if $u = v_{-1}$ or v_1, then $K_2(u \otimes u) = -u \otimes u$, so $K_N\bigl(u^{\otimes N}\bigr) = -(N-1)u^{\otimes N}$ and $Q_N\bigl(u^{\otimes N}\bigr) = -Nu^{\otimes N}$.

To investigate the largest eigenvalues of K_N and Q_N, we make the following definitions. For even N, let

$$\begin{aligned}
\mathcal{J}_{0,N} &= \operatorname{span}\bigl\{e_{j_1} \otimes e_{j_2} \otimes \cdots \otimes e_{j_N} \bigm| \text{each of } e_1, e_2, \text{ and } e_3 \text{ appears in an} \\
&\qquad\qquad \text{even number of positions}\bigr\}, \\
\mathcal{J}_{1,N} &= \operatorname{span}\bigl\{e_{j_1} \otimes e_{j_2} \otimes \cdots \otimes e_{j_N} \bigm| \text{each of } e_2 \text{ and } e_3 \text{ appears in an odd number} \\
&\qquad\qquad \text{of positions, and } e_1 \text{ appears in an even number of positions}\bigr\}, \\
\mathcal{J}_{2,N} &= \operatorname{span}\bigl\{e_{j_1} \otimes e_{j_2} \otimes \cdots \otimes e_{j_N} \bigm| \text{each of } e_1 \text{ and } e_3 \text{ appears in an odd number} \\
&\qquad\qquad \text{of positions, and } e_2 \text{ appears in an even number of positions}\bigr\}, \\
\mathcal{J}_{3,N} &= \operatorname{span}\bigl\{e_{j_1} \otimes e_{j_2} \otimes \cdots \otimes e_{j_N} \bigm| \text{each of } e_1 \text{ and } e_2 \text{ appears in an odd number} \\
&\qquad\qquad \text{of positions, and } e_3 \text{ appears in an even number of positions}\bigr\}.
\end{aligned}$$

For odd N, make the same definitions, except write "odd" for "even" and "even" for "odd."

Clearly, $\mathcal{V}^{\otimes N} = \mathcal{J}_{0,N} \oplus \mathcal{J}_{1,N} \oplus \mathcal{J}_{2,N} \oplus \mathcal{J}_{3,N}$. Each of these four subspaces is easily seen to be invariant under an operator of the form $X^{(i)}X^{(j)}$, where X is one of R, S, and T, and hence to be invariant under K_N and Q_N. Furthermore, by Theorems 5.4 and 5.8, we have $\mathcal{M}_{0,N} \subseteq \mathcal{J}_{0,N}$.

For a function $h : \{1, 2, \ldots, N\} \to \{1, 2, 3\}$, define

$$e_h \equiv e_{h(1)} \otimes e_{h(2)} \otimes \cdots \otimes e_{h(N)} \in \mathcal{V}^{\otimes N},$$

and

$$(-1)^h \equiv (-1)^{\#(h)},$$

where

$$\#(h) \equiv \text{ number of pairs } (i, j) \text{ with } i < j \text{ and } h(i) > h(j).$$

Define $\omega_{j,N} \in \mathcal{J}_{j,N}$, for $j = 0, 1, 2, 3$, by

$$\omega_{j,N} = \sum_{\substack{h \\ e_h \in \mathcal{J}_{j,N}}} (-1)^h e_h,$$

where the sum is taken over all functions $h : \{1, 2, \ldots, N\} \to \{1, 2, 3\}$ with e_h in the given subspace.

It is easily seen that $\Pi(\omega_{0,N}) = \omega_{0,N}$, where Π is the rotation operator defined at the end of §3. Furthermore, $\omega_{0,1} = 0$, $\omega_{0,2} = \phi$, $\omega_{0,3} = \Delta$,

$$\omega_{0,N} = \omega_{1,N-1} \otimes e_1 + (-1)^N \omega_{2,N-1} \otimes e_2 + \omega_{3,N-1} \otimes e_3,$$

and

$$\omega_{0,N} = \omega_{0,N-2} \otimes \phi + (-1)^{N-1} \Psi(\omega_{0,N-1}).$$

It follows immediately that with Λ and Ψ as defined in §5,

$$\Lambda(\omega_{0,N}) = (-1)^{N-1} \omega_{0,N-1}$$

and

$$\omega_{0,N} = \omega_{0,N-2} \otimes \phi + (-1)^{N-1} \Psi(\omega_{0,N-1});$$

this second equation shows $\omega_{0,N} \in \mathcal{M}_{0,N}$ for all N. More generally, it follows by similar reasoning that for $j = 0, 1, 2, 3$,

$$\Lambda(\omega_{j,N}) = (-1)^{j+N-1} \omega_{j,N-1}$$

and

$$\omega_{j,N} = \omega_{j,N-2} \otimes \phi + (-1)^{j+N-1} \Psi(\omega_{j,N-1}).$$

It is easily seen from the definitions that each of $\omega_{0,N}$, $\omega_{1,N}$, $\omega_{2,N}$, and $\omega_{3,N}$ is scalar plus skew-symmetric in each two adjacent tensor positions. It follows immediately that $\langle \Theta^{(k,k+1)}(\omega_{j,N}), \omega_{j,N} \rangle \geq \langle \omega_{j,N}, \omega_{j,N} \rangle$ for any k and for $j = 0, 1, 2, 3$. We improve this result in Lemma 6.6 below.

The vector $\omega_{0,N}$ has the following uniqueness property.

LEMMA 6.3. *Let $N \geq 2$ be even. Let $y \in \mathcal{J}_{0,N}$ be scalar plus skew-symmetric in each of the two adjacent tensor positions. Then y is a scalar multiple of $\omega_{0,N}$.*

Proof. Note that $\langle \omega_{0,N}, e_1^{\otimes N} \rangle = 1$, so there is a complex number λ such that if $z = y + \lambda \omega_{0,N}$, then $\langle z, e_1^{\otimes N} \rangle = 0$. We wish to show that $z = 0$. We proceed by induction on even N, the induction hypothesis being that if $z \in \mathcal{J}_{0,N}$ is such that z satisfies the hypothesis of y in Lemma 6.3, and if furthermore z is perpendicular to

$e_1^{\otimes N}$, then $z = 0$. For $N = 2$, the hypothesis is clearly true. Assuming it is true for $N-2$, with $N \geq 4$, write

$$z = \phi \otimes u_0 + (e_1 \otimes e_2 - e_2 \otimes e_1) \otimes u_3 + (e_2 \otimes e_3 - e_3 \otimes e_2) \otimes u_1 + (e_3 \otimes e_1 - e_1 \otimes e_3) \otimes u_2,$$

for some vectors $u_0, u_1, u_2, u_3 \in \mathcal{V}^{\otimes N-2}$. Note that u_0 satisfies the same properties as does z, so by the induction hypothesis, $u_0 = 0$. This means that z is skew-symmetric in tensor positions 1 and 2. Similarly, z is skew-symmetric in tensor positions k and $k+1$ for each k. Since $N \geq 4$, and z must be built out of only the three vectors e_1, e_2, and e_3, this implies that $z = 0$. This completes the induction step and proves the lemma. □

PROPOSITION 6.4. *Let N be even. Then*

$$\omega_{0,N} = \sum_{\substack{\sigma \in S_N \\ \sigma \text{ canonical}}} (\text{sgn } \sigma)\, \Omega_\sigma \left(\phi^{\otimes N/2}\right),$$

with canonical defined as in the remark following Theorem 5.4. *Hence* $\omega_{0,N} \in \mathcal{M}_{0,N}$.

Proof. Let

$$y = \sum_{\substack{\sigma \in S_N \\ \sigma \text{ canonical}}} (\text{sgn } \sigma)\, \Omega_\sigma \left(\phi^{\otimes N/2}\right).$$

Then y satisfies the hypothesis of Lemma 6.3, so that y is a scalar multiple of $\omega_{0,N}$. To show the scalar is 1, note that $\langle \omega_{0,N}, e_1^{\otimes N} \rangle = 1$, while

$$\langle y, e_1^{\otimes N} \rangle = \sum_{\substack{\sigma \in S_N \\ \sigma \text{ canonical}}} (\text{sgn } \sigma),$$

and this last expression is easily seen to be 1 by induction. □

Remarks. (1) The analogous statements to Lemma 6.3 for $\omega_{j,N}$, for $j = 0, 1, 2, 3$, and for N even or odd, are also true.

(2) In addition to Proposition 6.4, we also have the following. For odd N,

$$\omega_{0,N} = \sum_{\substack{\sigma \in S_N \\ \sigma \text{ canonical}}} (\text{sgn } \sigma)\, \Omega_\sigma \left(\phi^{\otimes (N-3)/2} \otimes \Delta\right),$$

with canonical as defined in the remark following Theorem 5.8. Also, for $j = 1, 2, 3$,

$$\omega_{j,N} = \sum_{\sigma} (\text{sgn } \sigma)\, \Omega_\sigma \left(\phi^{\otimes (N-1)/2} \otimes e_j\right),$$

with the sum taken over all $\sigma \in S_N$ with $\sigma(2i-1) < \sigma(2i)$ for $i = 1, 2, \ldots, (N-1)/2$, and with $\sigma(2) < \sigma(4) < \cdots < \sigma(N-1)$. For even N, and for $j = 1, 2, 3$,

$$\omega_{j,N} = \sum_{\sigma} (\text{sgn } \sigma)\, \Omega_\sigma \left(\phi^{\otimes (N-2)/2} \otimes (e_k \otimes e_l \ - \ e_l \otimes e_k)\right),$$

where $k < l$ and $\{j, k, l\} = \{1, 2, 3\}$ as sets, and with the sum taken over all $\sigma \in S_N$ with $\sigma(2i-1) < \sigma(2i)$ for $i = 1, 2, \ldots, N/2$, and with $\sigma(2) < \sigma(4) < \cdots < \sigma(N-2)$.

LEMMA 6.5. *For even N, $\|\omega_{0,N}\|^2 = \frac{1}{4}(3^N + 3)$ and $\|\omega_{1,N}\|^2 = \|\omega_{2,N}\|^2 = \|\omega_{3,N}\|^2 = \frac{1}{4}(3^N - 1)$. For odd N, $\|\omega_{0,N}\|^2 = \frac{1}{4}(3^N - 3)$ and $\|\omega_{1,N}\|^2 = \|\omega_{2,N}\|^2 = \|\omega_{3,N}\|^2 = \frac{1}{4}(3^N + 1)$.*

Proof. Let N be even. We have that

$$\|\omega_{0,N}\|^2 = \sum_{\substack{h \\ e_h \in \mathcal{J}_{0,N}}} 1.$$

To evaluate this number, note that it is equal to the number of ways of partitioning the integers $\{1, 2, \ldots, N\}$ into three disjoint subsets, each of which has even cardinality. Hence

$$\|\omega_{0,N}\|^2 = \sum_{\substack{N_1+N_2 \le N \\ N_1, N_2 \text{ even}}} \binom{N}{N_1}\binom{N-N_1}{N_2}.$$

This last expression is simply the sum of the coefficients of all terms in the expansion of the polynomial $(a+b+c)^N$ corresponding to even powers of each of a, b, and c, and is thus equal to

$$\frac{1}{4}\big((1+1+1)^N + (-1+1+1)^N + (1-1+1)^N + (1+1-1)^N\big) = \frac{1}{4}\left(3^N + 3\right).$$

The other norms may be similarly evaluated. □

LEMMA 6.6. *For even N,*

$$\frac{\langle K_N(\omega_{0,N}), \omega_{0,N}\rangle}{\langle \omega_{0,N}, \omega_{0,N}\rangle} = (N-1)\left(\frac{4}{3} + \frac{8}{3^N+3}\right),$$
$$\frac{\langle Q_N(\omega_{0,N}), \omega_{0,N}\rangle}{\langle \omega_{0,N}, \omega_{0,N}\rangle} = N\left(\frac{4}{3} + \frac{8}{3^N+3}\right),$$

and

$$\frac{\langle K_N(\omega_{1,N}), \omega_{1,N}\rangle}{\langle \omega_{1,N}, \omega_{1,N}\rangle} = \frac{\langle K_N(\omega_{2,N}), \omega_{2,N}\rangle}{\langle \omega_{2,N}, \omega_{2,N}\rangle} = \frac{\langle K_N(\omega_{3,N}), \omega_{3,N}\rangle}{\langle \omega_{3,N}, \omega_{3,N}\rangle}$$
$$= (N-1)\left(\frac{4}{3} - \frac{8}{3^{N+1}-3}\right),$$

while for odd N,

$$\frac{\langle K_N(\omega_{0,N}), \omega_{0,N}\rangle}{\langle \omega_{0,N}, \omega_{0,N}\rangle} = (N-1)\left(\frac{4}{3} - \frac{8}{3^N-3}\right),$$
$$\frac{\langle Q_N(\omega_{0,N}), \omega_{0,N}\rangle}{\langle \omega_{0,N}, \omega_{0,N}\rangle} = N\left(\frac{4}{3} - \frac{8}{3^N-3}\right),$$

and

$$\frac{\langle K_N(\omega_{1,N}), \omega_{1,N}\rangle}{\langle \omega_{1,N}, \omega_{1,N}\rangle} = \frac{\langle K_N(\omega_{2,N}), \omega_{2,N}\rangle}{\langle \omega_{2,N}, \omega_{2,N}\rangle} = \frac{\langle K_N(\omega_{3,N}), \omega_{3,N}\rangle}{\langle \omega_{3,N}, \omega_{3,N}\rangle}$$
$$= (N-1)\left(\frac{4}{3} + \frac{8}{3^{N+1}+3}\right).$$

Proof. Recall that

$$\omega_{0,N} = \omega_{0,N-2} \otimes \phi + (-1)^{N-1}\,\Psi\left(\omega_{0,N-1}\right).$$

Now, $\omega_{0,N-2}\otimes\phi$ is symmetric in tensor positions $N-1$ and N, while $(-1)^{N-1}\,\Psi\,(\omega_{0,N-1})$ is skew-symmetric in these tensor positions. Hence $\Theta_N^{(N-1,N)}\,(\omega_{0,N}) = 2\,\omega_{0,N-2}\otimes\phi + (-1)^{N-1}\,\Psi\,(\omega_{0,N-1}) = \omega_{0,N} + \omega_{0,N-2}\otimes\phi$. It follows that

$$\begin{aligned}\langle\Theta_N^{(1,2)}(\omega_{0,N}),\,\omega_{0,N}\rangle &= \|\omega_{0,N}\|^2 + \langle\omega_{0,N},\,\omega_{0,N-2}\otimes\phi\rangle \\ &= \|\omega_{0,N}\|^2 + \|\omega_{0,N-2}\|^2\|\phi\|^2 \\ &= \|\omega_{0,N}\|^2 + 3\,\|\omega_{0,N-2}\|^2.\end{aligned}$$

By symmetry, the same result holds when $(N-1,\,N)$ is replaced by $(k,k+1)$, so

$$\langle K_N(\omega_{0,N}),\omega_{0,N}\rangle = (N-1)\,(\|\omega_{0,N}\|^2 + 3\|\omega_{0,N-2}\|^2)\,.$$

The results for $K_N(\omega_{0,N})$ now follow from Lemma 6.5 and simple algebraic manipulation. The results for $Q_N(\omega_{0,N})$ follow similarly, using the fact that $\Pi(\omega_{0,N}) = \omega_{0,N}$. The results for $\omega_{1,N}$, $\omega_{2,N}$, and $\omega_{3,N}$ are proved similarly. □

THEOREM 6.7. *The operators $K_N|_{\mathcal{M}_{0,N}}$, $Q_N|_{\mathcal{M}_{0,N}}$, $K_N|_{\mathcal{J}_{1,N}}$, $K_N|_{\mathcal{J}_{2,N}}$, and $K_N|_{\mathcal{J}_{3,N}}$ have eigenvalues at least as large as the values, respectively, of*

$$\frac{\langle K_N(\omega_{0,N}),\omega_{0,N}\rangle}{\langle\omega_{0,N},\omega_{0,N}\rangle},\quad \frac{\langle Q_N(\omega_{0,N}),\omega_{0,N}\rangle}{\langle\omega_{0,N},\omega_{0,N}\rangle},\quad \frac{\langle K_N(\omega_{1,N}),\omega_{1,N}\rangle}{\langle\omega_{1,N},\omega_{1,N}\rangle},$$
$$\frac{\langle K_N(\omega_{2,N}),\omega_{2,N}\rangle}{\langle\omega_{2,N},\omega_{2,N}\rangle},\quad \textit{and}\quad \frac{\langle K_N(\omega_{3,N}),\omega_{3,N}\rangle}{\langle\omega_{3,N},\omega_{3,N}\rangle}$$

as given in Lemma 6.6. In particular, for any N, K_N has an eigenvalue larger than $\frac{4}{3}(N-1)$, and has at least four eigenvalues (counting multiplicity) asymptotically larger than or equal to $\frac{4}{3}N$. Also, Q_N has an eigenvalue asymptotically larger than or equal to $\frac{4}{3}N$.

Proof. Since the operators K_N and Q_N are Hermitian, this is immediate. □

The methods of this section also allow us to prove the following generalization of Corollary 4.2. It is previously known and, in fact, follows from the representation theory of $SU(2)$.

PROPOSITION 6.8. *The subspace $\mathcal{M}_{0,N}$ contains all eigenspaces of H_N and P_N of dimension 1 or 2.*

Proof. Let $\mathcal{E}$ be an eigenspace of H_N or P_N not contained in $\mathcal{M}_{0,N}$. Then by the decomposition of §4, there is a nonzero vector $w\in\mathcal{E}\cap\mathcal{M}_k^q$ for some $k\neq 0$ and $q=x$, y, or z. Now, it is easily seen that for any $k\neq 0$, $\mathcal{M}_k^q$ does not intersect $\mathcal{J}_{0,N}$. Indeed, $S_q^{(i)}\,(\mathcal{J}_{0,N})$ is orthogonal to $\mathcal{J}_{0,N}$, for each i, so no element of $\mathcal{J}_{0,N}$ can be an eigenvector of $\sum_{i=1}^N S_q^{(i)}$ corresponding to a nonzero eigenvalue. Hence $w\notin\mathcal{J}_{0,N}$. But this means that w has a nonzero projection onto some $\mathcal{J}_{j,N}$, $j>0$. Since the $\mathcal{J}_{j,N}$ are invariant subspaces of H_N and P_N, this projection is an element of $\mathcal{E}$. Then permuting the basis $\{e_1,e_2,e_3\}$ yields a nonzero element of $\mathcal{E}$ in each $\mathcal{J}_{j,N}$, $j=1,2,3$. Since the $\mathcal{J}_{j,N}$ are orthogonal, we must have $\dim\mathcal{E}\geq 3$, completing the proof. □

7. Other spin values. The vector space $\mathcal{V}$ and spin operators S_x, S_y, and S_z discussed in this paper correspond to atoms with spin 1. In general, spin values may be any element of $\{\frac{1}{2},1,\frac{3}{2},2,\ldots\}$. In this section, we discuss the spin operators for all spin values (see, e.g., [8]), and Haldane's more general conjecture [9], [10].

For spin $s\in\{\frac{1}{2},1,\frac{3}{2},2,\ldots\}$, the vector space required is $\mathcal{V}_s=\mathbb{C}^{2s+1}$ with orthonormal basis $\{v_{-s},v_{-s+1},\ldots,v_{s-1},v_s\}$. The operator S_z on $\mathcal{V}_s$ is defined by

$S_z(v_j) = jv_j$, extended by linearity. The operators S_x and S_y are defined, for $m, n \in \{-s, -s+1, \ldots, s\}$, by the relations

$$\langle v_m, S_x v_n \rangle = \frac{1}{2}\left[\sqrt{s(s+1) - n(n+1)}\,\delta_{m,n+1} + \sqrt{s(s+1) - m(m+1)}\,\delta_{m+1,n}\right]$$

and

$$\langle v_m, S_y v_n \rangle = \frac{1}{2i}\left[\sqrt{s(s+1) - n(n+1)}\,\delta_{m,n+1} - \sqrt{s(s+1) - m(m+1)}\,\delta_{m+1,n}\right]$$

extended by linearity, where δ is the Kronecker delta.

Once we have specified $s, \mathcal{V}_s, S_x, S_y$, and S_z, we may define the operators H_N and P_N on $\mathcal{V}_s^{\otimes N}$ exactly as before. These operators may be decomposed into operators on $\mathcal{M}_k^z$, $k \in \{-sN, -sN+1, \ldots, sN\}$, by exact analogy with §4 of this paper (invariance can be easily checked directly).

We may now state the general form of Haldane's conjecture [9], [10]. Let $\lambda_N^{0(s)}$ and $\lambda_N^{1(s)}$ be the smallest and second-smallest eigenvalues, respectively, of H_N (or P_N) with spin value s. Then Haldane's conjecture may be stated as saying that $\lim_{N\to\infty}(\lambda_N^{1(s)} - \lambda_N^{0(s)}) > 0$ if and only if s is an integer. (Again, there are other, inequivalent, mathematical formulations of the conjecture; see [1].) The mathematical proof or disproof of this conjecture would be of extreme interest in solid state physics. See [1] for a proof of the noninteger s statement, at least for P_N. The integer s statement, however, is considered to be more surprising, and remains unproven.

The following proposition is known and follows from the representation theory of $SU(2)$.

PROPOSITION 7.1. *If $s \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots\}$, and if N is odd, then each eigenspace of H_N and P_N has even dimension.*

Proof. As mentioned above, we decompose the space $\mathcal{V}_s^{\otimes N}$ into spaces $\mathcal{M}_k^z$, for $k = -sN, -sN+1, \ldots, sN$. Note that each of these values of k differs by $\frac{1}{2}$ from an integer, and in particular that k is never 0. It is easily checked that the mapping $v_m \mapsto v_{-m}$ induces a unitarily equivalence of $H_N|_{\mathcal{M}_k^z}$ with $H_N|_{\mathcal{M}_{-k}^z}$, and of $P_N|_{\mathcal{M}_k^z}$ with $P_N|_{\mathcal{M}_{-k}^z}$. Let $\mathcal{M}^+ = \bigcup_{k>0}\mathcal{M}_k^z$, and let $\mathcal{M}^- = \bigcup_{k<0}\mathcal{M}_k^z$. Then $\mathcal{V}_s^{\otimes N} = \mathcal{M}^+ \oplus \mathcal{M}^-$, and we have $H_N = H_N^+ \oplus H_N^-$ and $P_N = P_N^+ \oplus P_N^-$, where $H_N^+ = H_N|_{\mathcal{M}^+}$, etc. The result now follows from the observation that H_N^+ is unitarily equivalent to H_N^-, and P_N^+ is unitarily equivalent to P_N^-. □

Acknowledgments. The authors are grateful to R. L. Armstrong and W. J. L. Buyers for introducing us to these questions and to the editors and referees of this journal for providing a number of references and several useful suggestions.

REFERENCES

[1] I. AFFLECK AND E. H. LIEB, *A proof of part of Haldane's conjecture on spin chains*, Lett. Math. Phys., 12 (1986), pp. 57–69.

[2] I. AFFLECK, T. KENNEDY, E. H. LIEB, AND H. TASAKI, *Rigorous results on valence-bond ground states in antiferromagnets*, Phys. Rev. Lett., 59 (1987), pp. 799–802.

[3] ———, *Valence bond ground states in isotropic quantum antiferromagnets*, Comm. Math. Phys., 115 (1988), pp. 477–528.

[4] R. BOTET AND R. JULLIEN, *Ground-state properties of a spin-1 antiferromagnetic chain*, Phys. Rev. B., 27 (1983), pp. 613–615.

[5] W. J. L. Buyers, R. M. Morra, R. L. Armstrong, M. J. Hogan, P. Gerlach, and K. Hirakawa, *Experimental evidence for the Haldane gap in a spin-1, nearly isotropic, antiferromagnetic chain*, Phys. Rev. Lett., 56 (1986), pp. 371–374.

[6] K. Chang, Senior thesis, Princeton University, Princeton, NJ, 1987.

[7] K. Chang, I. Affleck, G. W. Hayden, and Z. G. Soos, *A study of the bilinear-biquadratic spin-1 antiferromagnetic chain using the valence-bond basis*, J. Phys.: Condensed Matter, 1 (1989), pp. 153–167.

[8] C. Cohen-Tannoudji, D. Diu, and F. Laloë, *Quantum Mechanics*, Vol. 1, Chap. VI, Hermann, Paris, 1977.

[9] F. D. M. Haldane, *Continuum dynamics of the 1-D Heisenberg antiferromagnet: Identification with the $O(3)$ nonlinear sigma model*, Phys. Lett., 93A (1983), pp. 464–468.

[10] ———, *Nonlinear field theory of large-spin Heisenberg antiferromagnets: Semiclassically quantized solitons of the one-dimensional easy-axis Neél state*, Phys. Rev. Lett., 50 (1980), pp. 1153–1156.

[11] M. P. Nightingale and H. W. J. Blöte, *Gap of the linear spin-1 Heisenberg antiferromagnet: A Monte Carlo calculation*, Phys. Rev. B, 33 (1986), pp. 659–661.

[12] M. Steiner, K. Kakurai, J. K. Kjems, D. Petitgrand, and R. Pynn, *Inelastic neutron scattering studies on 1-D near-Heisenberg antiferromagnets: A test of the Haldane conjecture*, J. Appl. Phys., 61 (1987), pp. 3953–3955.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 853–879, July 1993

PARALLEL SPARSE LU DECOMPOSITION ON A MESH NETWORK OF TRANSPUTERS*

A. FRANK VAN DER STAPPEN†‡, ROB H. BISSELING†§, AND JOHANNES G. G. VAN DE VORST†

Abstract. A parallel algorithm is presented for the LU decomposition of a general sparse matrix on a distributed-memory MIMD multiprocessor with a square mesh communication network. In the algorithm, matrix elements are assigned to processors according to the grid distribution. Each processor represents the nonzero elements of its part of the matrix by a local, ordered, two-dimensional linked-list data structure. The complexity of important operations on this data structure and on several others is analysed. At each step of the algorithm, a parallel search for a set of m compatible pivot elements is performed. The Markowitz counts of the pivot elements are close to minimum, to preserve the sparsity of the matrix. The pivot elements also satisfy a threshold criterion, to ensure numerical stability. The compatibility of the m pivots enables the simultaneous elimination of m pivot rows and m pivot columns in a rank-m update of the reduced matrix. Experimental results on a network of 400 transputers are presented for a set of test matrices from the Harwell–Boeing sparse matrix collection.

Key words. sparse matrices, LU decomposition, parallel algorithms, distributed-memory multiprocessor, transputers

AMS subject classifications. 65F05, 65F50, 65Y05

1. Introduction.

Sparse linear systems of equations need to be solved in many application areas, such as oil reservoir simulation, chemical plant modelling, and linear programming. A sparse linear system of equations has the form

$$Ax = b, \tag{1}$$

where A is an $n \times n$ sparse matrix, and x and b are vectors of length n. Vector b is given, and x is the unknown solution vector. In this paper we assume that the matrix A is nonsingular, sparse (i.e., cn of its n^2 elements have a nonzero value, with $c \ll n$), and general (i.e., A has an arbitrary, not necessarily symmetric pattern of nonzeros). The computing time and the amount of memory needed to solve (1) can be reduced greatly by exploiting the sparsity of A.

Several methods exist for solving the sparse system $Ax = b$ [12], [34]. One of these methods is based on LU decomposition [19], which is closely related to Gaussian elimination. LU decomposition produces an $n \times n$ unit lower triangular matrix L, an $n \times n$ upper triangular matrix U, and permutations π and ρ of $\{0, \ldots, n-1\}$ such that

$$A_{\pi_i, \rho_j} = (LU)_{ij} \quad \text{for all } i, j, \ 0 \le i, j < n. \tag{2}$$

Permutations π and ρ appear in this equation because rows and columns may have to be permuted during the LU decomposition to preserve sparsity and ensure numerical stability.

* Received by the editors March 5, 1991; accepted for publication (in revised form) December 18, 1991.

† Koninklijke/Shell-Laboratorium, Amsterdam, P.O. Box 3003, 1003 AA Amsterdam, the Netherlands.

‡ Present address, Department of Computer Science, Utrecht University, P.O. Box 80089, 3508 TB Utrecht, the Netherlands (frankst@cs.ruu.nl).

§ Present address, Department of Mathematics, Utrecht University, P.O. Box 80010, 3508 TA Utrecht, the Netherlands (bisseling@math.ruu.nl).

The system $Ax = b$ can be solved in five stages:

1. Decompose A to obtain matrices L, U and permutations π, ρ satisfying (2).
2. Permute b according to $d_i = b_{\pi_i}$, $0 \le i < n$, to obtain a vector d.
3. Solve $Ly = d$ to obtain a vector y. This unit lower triangular system of equations is solved by forward substitution.
4. Solve $Uz = y$ to obtain a vector z. This upper triangular system of equations is solved by back-substitution.
5. Permute z according to $x_{\rho_j} = z_j$, $0 \le j < n$, to obtain the solution vector x.

This paper deals exclusively with the first stage: we present an algorithm for the LU decomposition of a sparse matrix on a distributed-memory parallel computer. A suitable algorithm for the parallel solution of sparse triangular systems can be derived from the parallel dense triangular system solving algorithm in [2]. For a symmetric positive definite matrix A, it is more efficient to use Cholesky factorisation [19] instead of LU decomposition; an extensive review of parallel sparse Cholesky factorisation algorithms is given in [21].

Sequential sparse LU decompositions usually consist of many steps, each of which contains a search of the reduced matrix for one pivot element, followed by row and column permutations, and a rank-1 update of the reduced matrix. Sparsity can be preserved during the LU decomposition by choosing appropriate pivot elements. A heuristic pivot search strategy that achieves this aim is the Markowitz strategy [24] (see [14] for an experimental evaluation of its performance). The choice of a pivot element A_{ij} leads to the creation of at most $M_{ij} = (R_i - 1)(C_j - 1)$ new nonzeros, where R_i (C_j) is the number of nonzeros in row i (column j) of the reduced matrix. The upper bound M_{ij} is called the *Markowitz count* of A_{ij} [12, Chap. 7]. A pivot element that has the lowest Markowitz count, *mincount*, is chosen. Many variants of this basic Markowitz strategy exist. Zlatev [33] limits the search for pivot elements to the sparsest three rows, to prevent long searches; his experiments show that the total number of new nonzeros created, the *fill-in*, is not necessarily reduced by searching more rows.

Numerical stability must be maintained during LU decomposition to obtain an accurate solution. Sequential general-purpose programs for sparse LU decomposition such as MA28 [11] and Y12M [35] incorporate a mechanism to prevent small matrix elements from being chosen as a pivot. One variant is to accept only those pivot candidates A_{ij} that are nonzero and that satisfy

$$|A_{ij}| \ge u \cdot \max_l |A_{lj}|, \tag{3}$$

where u, $0 \le u \le 1$, is a threshold parameter [12, Chap. 9].

The aim of this paper is to present a parallel general-purpose sparse LU decomposition algorithm that includes features such as Markowitz pivoting to preserve sparsity and threshold pivoting to ensure numerical stability. The algorithm is suitable for distributed-memory message-passing multiprocessors such as transputer networks and hypercubes. The functionality of our parallel implementation is comparable to that of sequential programs such as MA28 and Y12M.

The potential parallelism in sparse LU decomposition is twofold: first, in dense LU decompositions the operations of each rank-1 update can be done in parallel; this parallelism is inherited by sparse algorithms. Second, the sparsity of the matrix allows for the parallel execution of some computations which in the dense case have to be performed in sequence. In the sparse case, several rank-1 updates of the matrix can be combined into one multiple-rank update with many potentially simultaneous operations, thereby avoiding synchronisation and idling of processors after each rank-1 update.

Both forms of parallelism have been exploited in shared-memory parallel sparse LU decomposition algorithms of Smart and White [29], Alaghband [1], Davis and Yew [8], [9], Gallivan, Sameh, and Zlatev [17], and others. In these algorithms the Markowitz strategy is modified to obtain a set S of pivot elements that can be handled simultaneously. To make this set as large as possible, pivot elements with a Markowitz count higher than *mincount* are accepted. Calahan [5] was the first to exploit the fact that two pivot elements A_{ij} and A_{kl} can be handled simultaneously if

$$A_{il} = A_{kj} = 0. \tag{4}$$

Such pivots are called *compatible* [1] or *independent* [29]. For a detailed discussion of compatibility, see [9].

Smart and White [29] investigate the parallel time complexity of sparse LU decomposition for an unlimited number of processors, by using task graph depths as a complexity measure. They present an algorithm in which the pivot set S contains compatible diagonal elements with a Markowitz count between *mincount* and *mincount* $+ a$, where a is an input parameter. The set S is constructed by starting with the empty set and successively adding new compatible pivot elements in order of increasing Markowitz count. For a 1000×1000 tridiagonal matrix, the algorithm with $a = 2$ leads to a task graph depth of 27, which is close to the theoretical minimum of 23; in this case, the basic Markowitz strategy leads to a much higher depth of 1998. For most of the examined electronic circuit matrices, however, the improvement in depth over the Markowitz strategy was only about 50 percent.

Alaghband [1] presents an algorithm which generates candidate pivot sets and then chooses a pivot set S of maximum size; ties between sets are decided according to the minimum total Markowitz sum. The pivot search uses an $n \times n$ table that represents the mutual compatibility of diagonal pivot candidates. Pivot elements with a Markowitz count higher than a user-specified value or with a numerical absolute value lower than a user-specified threshold are discarded from S. Experimental results on the Denelcor HEP shared-memory computer show a speedup of 4.8 on eight processors, for a 144×144 electronic circuit matrix with 616 nonzeros. In this example many pivots are handled in parallel: the matrix is decomposed in 10 steps, with pivot sets of sizes $m = 72, 25, 16, 11, 6, 5, 3, 2, 2, 1, 1$.

Davis and Yew [8], [9] present a shared-memory parallel algorithm and a program, D2, which has the full functionality of programs such as MA28 and Y12M. In this algorithm, the pivot set S contains compatible elements with a Markowitz count between *mincount* and $a \cdot$ *mincount*, where a is an input parameter ($a = 4$ in the experiments). All processors search for acceptable pivot candidates, and then try to add them to the current set S. If a candidate is compatible with all the elements of S, it is added to S. Conflicts between processors that simultaneously try to add a pivot are prevented by critical sections in the program. The processor that is the first to arrive at the entry of a critical section gains access to it. This implies that operating system factors influence the pivot choice; the program is therefore nondeterministic. Experiments on an Alliant FX/8 shared-memory computer show that D2 is a median 3.9 times faster on eight processors than on a single processor, and that the sequential version of D2 is 4.3 times faster than the sequential program MA28.

Gallivan, Sameh, and Zlatev [17] (see also [34, Chap. 10]) present three shared-memory parallel versions of the sequential program Y12M [35]: Y12M1, which is based on rank-1 updates; Y12M2, which is based on rank-m updates; and Y12M3, which exploits coarse-grain parallelism created by ordering the matrix A to an upper block-triangular form (see also [16] for alternative ordering techniques). Here, the enhanced parallelism of Y12M2 compared to Y12M1 is important. In the algorithm Y12M2, an ordered set S of m pivots is formed with the property that $A_{kj} = 0$ if the element A_{ij} precedes the element A_{kl} in the ordering of the set. This relaxation of the compatibility requirement (4) allows for the creation of larger pivot sets, at the expense of a more complicated matrix update. Experimental results on an Alliant FX/80 shared-memory parallel computer with eight processors show that Y12M1 is faster than Y12M2 for matrices that are relatively dense or become so during the LU decomposition, whereas Y12M2 is faster for matrices that are very sparse and remain so. The speedup of Y12M1 is between 2.4 and 5.4 and that of Y12M2 is between 2.3 and 5.0, for a set of 27 test matrices from the Harwell–Boeing sparse matrix collection [13].

Sadayappan and Rao [26] analyse the amount of communication in sparse LU decomposition on a distributed-memory parallel computer. They present the *fragmented distribution* which splits rows and columns into parts and distributes these parts over different processors; this is in contrast to the shared-memory algorithms above that treat rows or columns as basic indivisible units. Compared to a row/column-wrapped distribution, the fragmented distribution decreases the communication volume (i.e., the total length of the messages) of dense LU decomposition by a factor of Q, for Q^2 processors. Statistics for a number of circuit simulation matrices confirm that the communication volume for sparse matrices is reduced in the same way as for dense matrices, showing up to a five-fold decrease for $Q = 8$.

Skjellum [28] presents a distributed-memory parallel algorithm that is valid for a range of data distributions, including the grid distribution defined below. The generality of this algorithm enables the user to tune the granularity of the distribution to the characteristics of a particular computer architecture. In the current implementation, the partial row pivoting strategy is used, which gives one pivot element per step. Experimental results on a Symult s2010 for a 2500 × 2500 random sparse matrix with $c \approx 51$ nonzeros per row show a speedup of 9.7 on a 96 processor machine, compared to a 6 processor machine.

Our distributed-memory parallel LU decomposition algorithm for sparse matrices is based on an algorithm for dense matrices [3]. The dense algorithm allocates matrix elements to processors according to the *grid distribution* [32], defined by the mapping

$$(5) \qquad A_{ij} \longmapsto \text{processor}\,(i \bmod Q, j \bmod Q) \quad \text{for all } i, j,\ 0 \le i, j < n,$$

for Q^2 processors (s,t), $0 \le s,t < Q$. This distribution splits each row i into Q *row parts*, i.e., sets of the form $\{A_{ij} : 0 \le j < n \wedge j \bmod Q = t\}$, and it also splits each column into Q *column parts*. The grid distribution is also called *scattered square decomposition* [15] and *cyclic storage* [22]; it is similar to the fragmented distribution [26].

We choose the grid distribution (5) as the distribution scheme for sparse matrices because it has an optimal load balance and a low communication complexity for LU decomposition of dense matrices [3]. The optimal load balance in all steps of the dense LU decomposition algorithm implies that in the sparse algorithm all processors are responsible for an approximately equal number of (zero or nonzero) elements. If the statistical assumption holds that every element of the matrix has an equal probability of being nonzero, it follows that the nonzero elements are spread evenly over the processors. If this assumption does not hold because nonzeros cluster at certain places in the matrix, e.g., in the lower right-hand corner or in dense submatrices, then these nonzeros are scattered over many processors, still giving a good overall load balance. Memory use is also efficient, since the scattering effect makes it unlikely that one processor runs out of memory while the others still have much storage space available. (For a theoretical analysis of one-dimensional scattering applied to an irregular computational domain, see [25].)

Figure 1 shows four snapshots of the parallel LU decomposition of the 59×59 sparse matrix IMPCOL B from the Harwell–Boeing collection. At the start (a) of the LU decomposition, the matrix has 3169 zero elements and 312 nonzero elements, and at the end (d) it has 3053 zeros and 428 nonzeros. The zero elements are shown in yellow. The nonzero elements are assigned to the processors of a 2×2 mesh, according to the grid distribution. The elements of processor $(0,0)$ are shown in blue; those of $(0,1)$ in red; those of $(1,0)$ in green; and those of $(1,1)$ in purple. The matrix is shown at the start of step k of Algorithm 1, for (a) $k = 0$; (b) $k = 18$; (c) $k = 38$; (d) $k = 59$. The horizontal lines in (b) and (c) separate the matrix elements (i,j) with $i < k$ from those with $i \geq k$, and similarly for the vertical lines. The diagonal blocks of (d) are formed during the steps of the decomposition. Each block contains the permuted pivot elements of one step. These elements are found in a search for at most $m_{\max} = 10$ compatible pivot elements. The block sizes are $m = 9, 4, 5, 8, 7, 5, 6, 6,$ and nine times 1.

An alternative to the matrix-independent grid distribution is a distribution that exploits knowledge of the sparsity pattern of the matrix to obtain an optimal load balance. Unfortunately, at every step of the LU decomposition of a general sparse matrix the sparsity pattern changes due to permutations and fill-in that are unpredictable, because they depend on the pivot choice and hence on the numerical values of the matrix elements. Adjustment to the changing sparsity pattern would necessitate frequent redistribution of the matrix, which is undesirable. For this reason, we do not adopt this approach and instead we rely on statistical expectations.

The low communication complexity of the dense LU decomposition algorithm implies a low communication complexity for the sparse algorithm, provided that the same statistical assumption as above holds. This can be seen as follows. In the dense case, the broadcast of a row of length n is done by simultaneously broadcasting Q row parts of length n/Q to Q processors each, with a time complexity of $\mathcal{O}(n/Q+Q)$. (The expressions $\mathcal{O}(x)$, $\Omega(x)$, and $\Theta(x)$ denote, respectively, at most, at least, and equal to a constant times x.) Similarly, in the sparse case, the broadcast of a row of length c is done by simultaneously broadcasting Q row parts of length c/Q to Q processors each, with a time complexity of $\mathcal{O}(c/Q + Q)$. This communication complexity is less than, for instance, the $\mathcal{O}(c + Q)$ complexity of a row broadcast for the row-wrapped distribution on a square mesh of processors. (The gain can be significant even for small values of c, where Q dominates c/Q, because in our algorithm m row broadcasts can be combined into one multiple-row broadcast of complexity $\mathcal{O}(mc/Q + Q)$.) Sadayappan and Rao [26] observed a similar reduction in communication.

The pivot search strategy of a distributed-memory parallel algorithm must be kept simple, to avoid excessive communication between searching processors. The chosen data distribution strongly influences the choice of the search heuristic. In our algorithm, each column $(*,t)$ of Q processors is responsible for about n/Q matrix columns. Each processor column searches a few of its sparsest matrix columns for numerically stable pivots with low Markowitz counts. A pivot set S of size m is then constructed by starting with the empty set and adding new compatible pivot candidates in order of increasing Markowitz count, similar to the algorithm of Smart and White [29]. This is done by a parallel algorithm which uses the distributed compatibility information supplied by the distributed matrix A. The pivot set is constructed by a pipeline of Q processors, and then broadcast to all Q^2 processors. After the pivot search, the matrix is permuted according to S and then modified by a rank-m update.

An important issue in sparse matrix computations is the choice of a data structure. In this paper, we analyse the theoretical time complexity of important parallel computations on several data structures. For reasons of simplicity, we decide to represent the nonzero elements of each processor in a local, ordered, two-dimensional linked-list data structure. This data structure has been introduced by Knuth [23] for sequential sparse matrix computations; it has been used by Alaghband [1] and Skjellum [28] for parallel computations. (Davis and Yew [8], [9] use two one-dimensional doubly linked-list structures, one for rows and one for columns. Each entry in such a structure represents a block of nonzeros.)

The remainder of this paper is organised as follows. In §2 we present the details of the parallel sparse LU decomposition algorithm. In §3 we review the possible choices of a local data structure for the local part of the sparse matrix and we compare their time complexity and memory use. In §4 we present the results of numerical experiments on square meshes of up to 400 transputers, for a test set of eleven matrices from the Harwell–Boeing sparse matrix collection [13]. In §5 we draw the conclusions.

2. Parallel algorithm.

In this section we present an algorithm for the parallel LU decomposition of sparse matrices on a square processor mesh with distributed memory. The algorithm consists of a number of steps, each of which has three phases: a pivot search, row and column permutations, and an update of the reduced matrix and its row and column nonzero counts.

2.1. Outline of the parallel algorithm.

We may choose to distribute a matrix over the processors and then find a local representation for each matrix part. On the other hand, we may choose to find a representation for the entire matrix and then distribute this representation. We decide to distribute the matrix first. This choice is justified by the following arguments:

- If distribution is accomplished first, the representation details of a matrix are always local. This locality eases the implementation of operations that change the sparsity pattern of the matrix. As an example, the insertion or deletion of an element in a local representation is an operation performed by a single processor.
- If representation is accomplished first, both the matrix itself and its representation details have to be distributed. This dual requirement has the consequence that operations that change the sparsity pattern of the matrix are always of a global nature. Hence, several processors have to cooperate to perform such operations. This cooperation makes these operations unnecessarily inefficient and difficult to implement.

Matrices are distributed over a square mesh of Q^2 processors. Each processor is identified by Cartesian coordinates (s,t), with $0 \leq s,t < Q$; in what follows we shall omit these bounds on s and t for the sake of brevity. We define $grid(s,t)$ as the set of index pairs of an $n \times n$ matrix assigned to processor (s,t) according to the grid distribution (5),

$$grid(s,t) = \{\ (i,j)\ :\ 0 \leq i,j < n \wedge i \bmod Q = s \wedge j \bmod Q = t\}. \tag{6}$$

We introduce an $n \times n$ matrix variable X, and distribute it according to the grid distribution. The permutation variables π and ρ of (2) are replicated and distributed: it turns out to be convenient to maintain a copy of π_i in all processors ($i \bmod Q$, $*$), and a copy of ρ_j in all processors ($*$, $j \bmod Q$). Initially $X = A$ and $\pi = \rho = \text{id}$, the identity permutation. At the end of the computation the matrix variable X contains the factor L in its strictly lower triangular part and U in its upper triangular part, and the permutation variables π and ρ contain their final values. Row nonzero counts are maintained in a vector variable R which is replicated and distributed in the same manner as π. The component R_i equals the number of nonzero elements of row i in the *reduced matrix*, which is defined as the $(n-k) \times (n-k)$ submatrix of elements $X_{ij}, k \leq i, j < n$, at the start of step k of the algorithm below. (For the purpose of explanation, we assume throughout this paper that there are no accidental zeros in the computations. Therefore, the number of nonzeros in a row equals the number of entries in the sparse representation of that row.) Similarly, column nonzero counts are maintained in a vector variable C which is replicated and distributed in the same manner as ρ. The Appendix presents a formal description of the aim of the algorithm and of the relation between X, π, ρ, R, and C that is maintained throughout the algorithm. An outline of the algorithm follows.

ALGORITHM 1 (parallel sparse LU decomposition).

```
$X := A$;
$\pi :=$ id;
$\rho :=$ id;
initialise $R$ and $C$;
$k := 0$;
**while** $k < n$ **do begin**
    find pivot set $S = \{(i_r, j_r) : 0 \leq r < m\}$;
    permute rows $i \in \{i : k \leq i < k+m\} \cup \{i_r : 0 \leq r < m\}$;
    permute corresponding $\pi_i$ and $R_i$;
    permute columns $j \in \{j : k \leq j < k+m\} \cup \{j_r : 0 \leq r < m\}$;
    permute corresponding $\rho_j$ and $C_j$;
    update matrix elements $\{X_{ij} : k+m \leq i < n \wedge k \leq j < k+m\}$;
    update matrix elements $\{X_{ij} : k+m \leq i, j < n\}$;
    update row nonzero counts $\{R_i : k+m \leq i < n\}$;
    update column nonzero counts $\{C_j : k+m \leq j < n\}$;
    $k := k+m$
**end.**
```

In the notation of this outline, it is implied that each processor (s,t) performs its part of the computations on its own data. The details of the separate parts of the algorithm are presented in §§2.2–2.4. The algorithm is illustrated in Fig. 1.

2.2. Parallel pivot search. A simple and effective pivot search strategy is to choose pivot elements from a limited number of the sparsest rows or columns; see [33]. Since we choose the column-oriented stability criterion (3), it is most convenient to search columns. The pivot search consists of three parts: searching columns to find candidate pivot elements; determining mutual compatibility of candidates; and constructing a pivot set of mutually compatible elements.

2.2.1. Search for candidates. In the first part of the pivot search, columns are inspected in parallel. Processor (s,t) has one column part of each column j with $k \leq j < n$ and $j \bmod Q = t$, and it participates in the search of *ncol* of these columns that have lowest nonzero counts C_j. Here, *ncol* is an input parameter. (If less than *ncol* columns remain, these are searched.) The set of columns to be searched by processor (s, t) is denoted by $SearchCols(t)$; this set is available in all processors $(*, t)$. Together, the processors $(*, t)$ search the complete columns of $SearchCols(t)$. A set $ColCandidates(t)$ is formed which includes one optimal pivot candidate per searched column; this set becomes available in all processors $(*, t)$. An outline of the program text for processor column $(*, t)$ follows.

ALGORITHM 2 (find candidate pivot elements).

```
  $ColCandidates(t) := \emptyset$;
  determine $SearchCols(t)$ from $C$;
  for all $j \in SearchCols(t)$ do begin
    find $r$, $k \le r < n$, such that $|X_{rj}| = \max\{|X_{lj}| : k \le l < n \wedge X_{lj} \ne 0\}$;
    $threshold := u \cdot |X_{rj}|$;
    find $i$, $k \le i < n$, such that
      $M_{ij} = \min\{M_{lj} : k \le l < n \wedge |X_{lj}| \ge threshold \wedge X_{lj} \ne 0\}$;
    $ColCandidates(t) := ColCandidates(t) \cup \{(i, j)\}$
  end
```

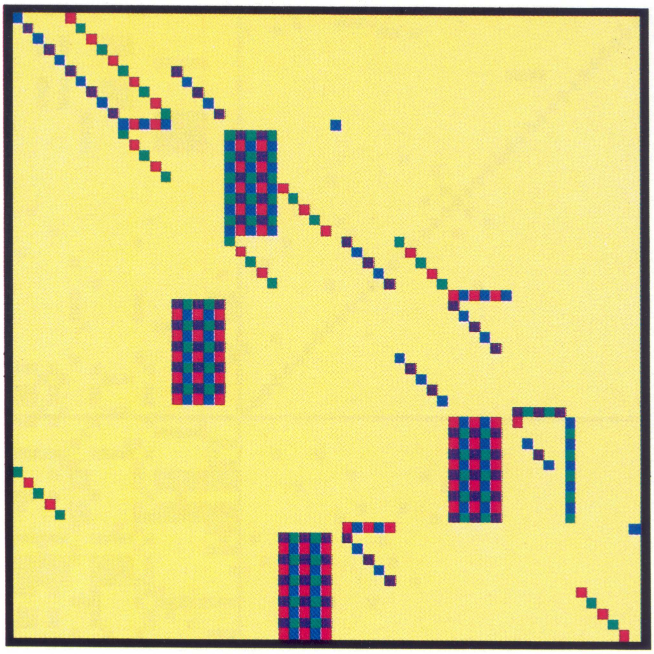
The statements of Algorithm 2 are implemented as follows. The set $SearchCols(t)$ is determined by using a local data structure for column nonzero counts (see [12, Chap. 9]). The index r of an element with maximum absolute value in column j is determined by first searching locally in processor (s, t), and then communicating and comparing these local maxima to obtain the index r of the global maximum. This maximum $|X_{rj}|$ is broadcast to all processors $(*, t)$. Markowitz counts M_{ij}, $(i, j) \in grid(s, t)$, are computed from locally available nonzero counts R_i, $i \bmod Q = s$, and C_j, $j \bmod Q = t$. The index i is determined in the same fashion as the index r.

After the sets $ColCandidates(t)$ are formed, they are collected into one set $Candidates = \bigcup_{t=0}^{Q-1} ColCandidates(t)$. The total number of candidates is $ncand = \min(Q \cdot ncol, n-k)$. The pivot candidates are sorted according to increasing Markowitz count, $Candidates = \{(i_r, j_r) : 0 \le r < ncand\}$, with $M_{i_r, j_r} \le M_{i_{r'}, j_{r'}}$ if $r < r'$. This ordering is used later on to give preference to candidates with low Markowitz counts. Now, candidates (i, j) that have an unacceptably high Markowitz count, $M_{ij} > a \cdot M_{i_0, j_0}$, are discarded.

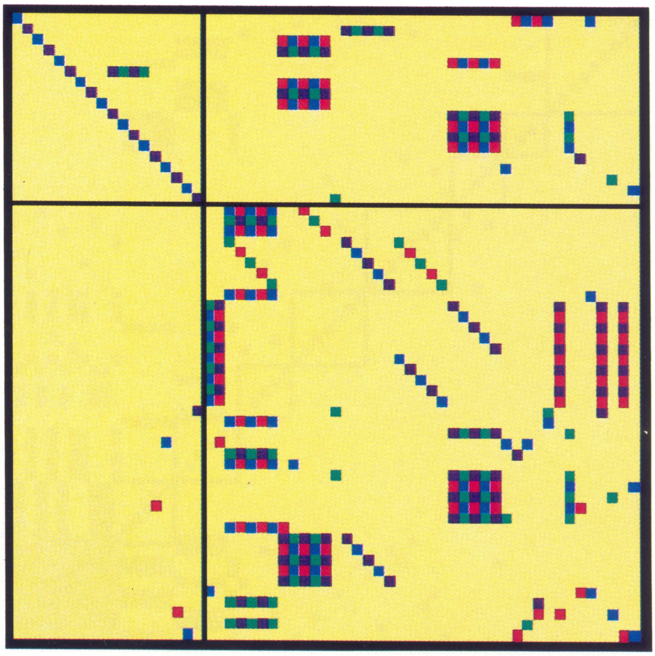
The set *Candidates* is sorted during its construction by using a pipeline of processors $(0, *)$ as follows. First, each processor $(0, t)$ sorts its own set $ColCandidates(t)$ by increasing Markowitz count. After that, processor $(0, t)$ inserts its candidates at the appropriate places in the ordered stream of candidates that passes by, going from processor $(0, t-1)$ to $(0, t+1)$. Processor $(0, 0)$ starts the pipeline. Processor $(0, Q-1)$ collects the ordered sequence into *Candidates* and broadcasts this set to all other processors.

2.2.2. Compatibility of candidates. In the second part of the pivot search, the compatibility of each pivot candidate with all other candidates is determined. The second part is separated from the third part, the construction of the pivot set, to avoid frequent synchronisation of processors, which would occur if these parts were combined. To determine compatibility, it is sufficient to inspect for each candidate (i, j) the column j, and to check whether there are nonzeros $X_{i'j}$ in rows i' that contain candidates $(i', j') \ne (i, j)$. Pivot candidate (i, j) is marked as incompatible with these candidates (i', j').

The work is distributed by letting each processor search its local part of the sparse

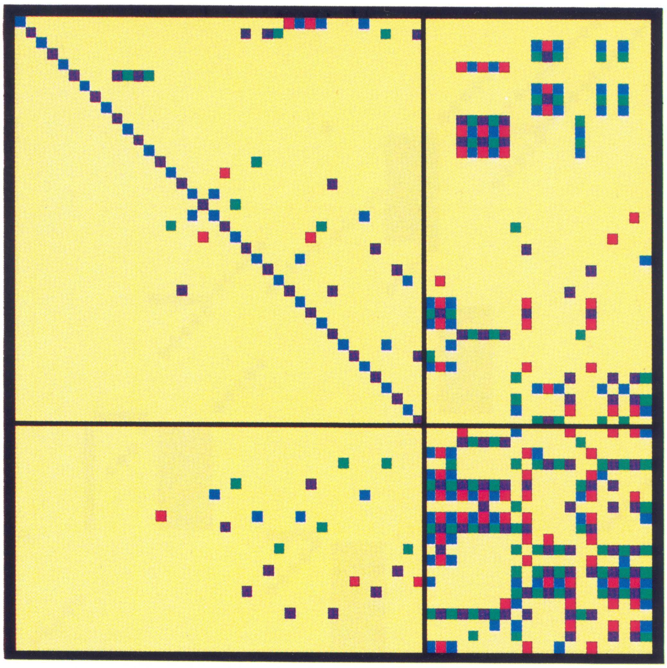


(a) $k = 0$

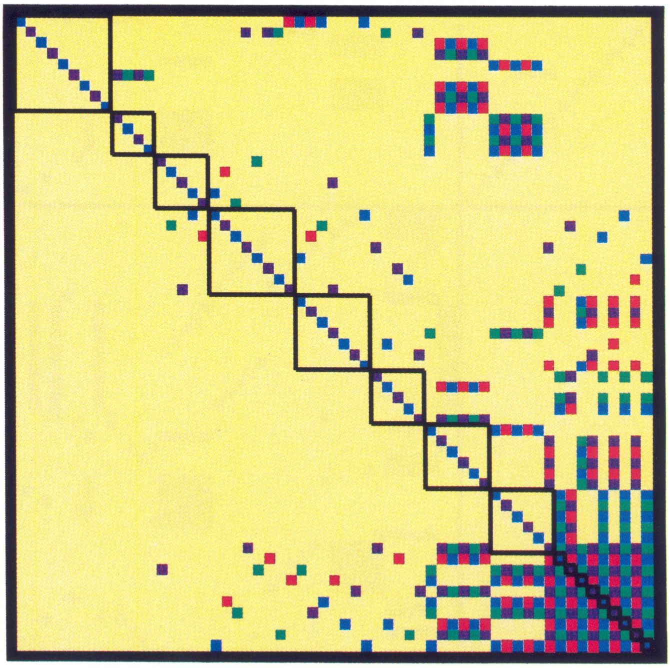


(b) $k = 18$

FIG. 1. *Snapshots of the sparse matrix* IMPCOL B *at step k of its parallel LU decomposition.*



(c) $k = 38$



(d) $k = 59$

matrix for nonzeros that make two candidates mutually incompatible. Processor (s,t) compares candidates from $\mathit{RowCandidates}(s) = \{(i,j) \in \mathit{Candidates} : i \bmod Q = s\}$ with candidates from $\mathit{ColCandidates}(t)$. The set $\mathit{RowCandidates}(s)$ is available in all processors $(s,*)$. In this way, the processors build a distributed incompatibility set. This set is often small, because the original matrix is sparse. Processor (s,t) stores its incompatibilities in a local set $\mathit{Incompatible}(s,t) = \{(i,j,i',j') : (i,j) \in \mathit{ColCandidates}(t) \wedge (i',j') \in \mathit{RowCandidates}(s) \wedge (i,j) \neq (i',j') \wedge X_{i'j} \neq 0\}$. An outline of the program text for processor (s,t) follows.

ALGORITHM 3 (determine compatibility of candidate pivot elements).

```
$\mathit{RowCandidates}(s) := \{(i,j) \in \mathit{Candidates} : i \bmod Q = s\}$;
$\mathit{Incompatible}(s,t) := \emptyset$;
for all $(i,j) \in \mathit{ColCandidates}(t)$ do
    for all $(i',j') \in \mathit{RowCandidates}(s)$ do
        if $(i,j) \neq (i',j') \wedge X_{i'j} \neq 0$ then
            $\mathit{Incompatible}(s,t) := \mathit{Incompatible}(s,t) \cup \{(i,j,i',j')\}$
```

2.2.3. Construction of the pivot set. In the third part of the pivot search, the pivot set S is constructed by starting with the empty set and successively adding new compatible pivot candidates (i_r, j_r) in order of increasing r. This procedure gives priority to candidates with lower Markowitz counts. The set S is constructed by a pipeline of processors which generates pivot elements, similar to a parallel sieve of Eratosthenes which generates prime numbers. Each processor is responsible for determining the inclusion in S of several pivot candidates. The length L of the pipeline can be chosen freely, between $1 \leq L \leq \min(\mathit{ncand}, Q^2)$. The choice $L = 1$ implies that all the work is done by one processor, giving an upper bound on the time complexity of $\mathcal{O}(Q^2 \cdot \mathit{ncol}^2)$, for a yield of $\mathcal{O}(\mathit{ncand}) = \mathcal{O}(Q \cdot \mathit{ncol})$ pivots. This is because in the worst case each candidate is checked for compatibility with all its predecessors. The best choice (for $\mathit{ncand} \leq Q^2$) is $L = \mathit{ncand}$, so that each processor determines the inclusion of exactly one candidate. The time complexity for this choice is $\mathcal{O}(Q \cdot \mathit{ncol})$. For practical reasons, we choose a length $L = Q$, and implement the pipeline in processor row $(0,*)$; see Algorithm 4 below. This can easily be modified, if desired. The time complexity for $L = Q$ has an upper bound of $\mathcal{O}(Q \cdot \mathit{ncol}^2)$, which is close to optimum for small ncol.

The responsibility for including a candidate (i_r, j_r) in the pivot set S or not is distributed evenly over the processors of the pipeline: each processor $(0,t)$ is responsible for at most ncol *local candidates*, i.e., the set $S(t)$ of candidates $(i_r, j_r), 0 \leq r < \mathit{ncand}$, with $\lfloor r/\mathit{ncol} \rfloor = t$. This set is needed only by processor $(0,t)$. To decide on inclusion in S, a processor $(0,t)$ needs the set $I(t)$ which contains the incompatibilities $(i_r, j_r, i_{r'}, j_{r'})$ and $(i_{r'}, j_{r'}, i_r, j_r)$ of each of the local candidates (i_r, j_r) with all the preceding candidates $(i_{r'}, j_{r'})$, $r' < r$. To provide this information, each $(i_r, j_r, i_{r'}, j_{r'}) \in \mathit{Incompatible}(s,t)$ is sent from processor (s,t) to processor $(0, \lfloor \max(r,r')/\mathit{ncol} \rfloor)$, before the pipeline starts operating.

The pipeline works as follows: processor $(0,t)$ receives a sequence of pivot elements from its neighbour $(0,t-1)$ and sends these elements to its neighbour $(0,t+1)$. If a pivot element from the sequence is incompatible with a local candidate, that local candidate is eliminated. After the pivot elements have passed, processor $(0,t)$ treats the sequence of remaining local candidates in a similar manner. The pipeline is started by processor $(0,0)$, and the m pivot elements of S are collected by processor $(0, Q-1)$ and then broadcast to all processors. An outline of the program text for processor $(0,t)$ follows.

ALGORITHM 4 (construct pivot set).

```
if $t = Q - 1$ then $S := \emptyset$;
$S(t) := \{(i_r, j_r) : 0 \le r < ncand \wedge \lfloor r/ncol \rfloor = t\}$;
while a pivot element $(i,j)$ is received from $(0, t-1)$ do begin
    if $t < Q - 1$ then send $(i,j)$ to $(0, t+1)$
                  else $S := S \cup \{(i,j)\}$;
    $S(t) := S(t) \setminus \{(i',j') : (i,j,i',j') \in I(t) \vee (i',j',i,j) \in I(t)\}$
end;
while $S(t) \ne \emptyset$ do begin
    $(i,j) := (i_l, j_l)$ with $l = \min\{r : (i_r, j_r) \in S(t)\}$;
    $S(t) := S(t) \setminus \{(i,j)\}$;
    if $t < Q - 1$ then send $(i,j)$ to $(0, t+1)$
                  else $S := S \cup \{(i,j)\}$;
    $S(t) := S(t) \setminus \{(i',j') : (i,j,i',j') \in I(t) \vee (i',j',i,j) \in I(t)\}$
end
```

2.3. Parallel permutations. Rows and columns of the matrix X can be permuted *implicitly* [28], [31] by using the permutations π and ρ to access the matrix indirectly, or *explicitly* [3], [6], [18] by moving rows and columns in the matrix. Chu and George [6] show that explicit permutation leads to a good load balance for parallel dense LU decomposition with a row-wrapped distribution. Numerical experiments of Geist and Romine [18] confirm that the gain in load balance more than offsets the incurred increase in communication time.

For dense LU decomposition with partial pivoting, the grid distribution with explicit permutation leads to an optimal load balance [3], irrespective of the choice of pivots. The load balance for implicit permutation, however, hinges on the randomising effect of the particular pivot sequence. For sparse LU decomposition with the grid distribution, explicit permutation guarantees that the row and column parts of the reduced matrix are evenly distributed over the processors. Therefore, the computational workload is well balanced if the assumption holds that the nonzeros of the matrix are evenly distributed over the row and column parts. Because of these considerations, we decide to permute rows and columns explicitly.

The aim of the row and column permutations in the matrix X is to create an $m \times m$ diagonal submatrix of elements $X_{ij}, k \le i,j < k+m$, with the m elements of the pivot set S on the diagonal; see Fig. 1(d). Because of the compatibility of the pivot elements this can be achieved, for instance, by moving the rows i_r, $0 \le r < m$, into position $k+r$ and moving the rows of $\{i : k \le i < k+m\} \setminus \{i_r : 0 \le r < m\}$ in some arbitrary order into the vacated positions $i_r \ge k+m$. The columns should be treated accordingly. Note that there is some freedom in determining the row permutation, particularly for large m; this may be exploited by a heuristic strategy which keeps row movements local, or even tries to avoid them (if $k \le i_r < k+m$).

The row permutation involves at most the rows of $\{i : k \le i < k+m\} \cup \{i_r : 0 \le r < m\}$; all the other rows remain in place. Note that the intersection of both sets may not be empty. The source and destination indices of the rows to be moved are stored in arrays *Src* and *Dest* of length *ndest*. Source processor (s,t) sends its part of row $Src(r)$ to destination processor $(Dest(r) \bmod Q, t)$, for each index r, $0 \le r < ndest$, with $Src(r) \bmod Q = s$. These communications are most efficiently implemented by operating $2Q$ pipelines in parallel, one upwards and one downwards for each processor column $(*, t)$. Data are injected into the appropriate pipeline by the source processor; they flow downstream until they are extracted by the destination processor. After this communication phase, processor (s, t) performs the following assignments.

ALGORITHM 5 (permute rows and row nonzero counts).

```
for all $r : 0 \leq r < ndest \wedge Dest(r) \bmod Q = s$ do begin
    for all $j : 0 \leq j < n \wedge j \bmod Q = t \wedge (X_{Src(r),j} \neq 0 \vee X_{Dest(r),j} \neq 0)$
        do $X_{Dest(r),j} := X_{Src(r),j}$;
    $\pi_{Dest(r)} := \pi_{Src(r)}$;
    $R_{Dest(r)} := R_{Src(r)}$
end
```

2.4. Parallel updates. An algorithm for processor (s, t), which updates the matrix, follows.

ALGORITHM 6 (update matrix).

```
broadcast $\{X_{jj} : k \leq j < k+m \wedge j \bmod Q = t\}$ from $(t,t)$ to $(*,t)$;
for all $i,j : k+m \leq i < n \wedge k \leq j < k+m \wedge (i,j) \in grid(s,t) \wedge X_{ij} \neq 0$
    do $X_{ij} := X_{ij}/X_{jj}$;
broadcast $\{X_{il} : k+m \leq i < n \wedge k \leq l < k+m \wedge (i,l) \in grid(s,t) \wedge$
    $X_{il} \neq 0\}$ from $(s,t)$ to $(s,*)$;
broadcast $\{X_{lj} : k \leq l < k+m \wedge k+m \leq j < n \wedge (l,j) \in grid(s,t) \wedge$
    $X_{lj} \neq 0\}$ from $(s,t)$ to $(*,t)$;
for all $i,j,l : k+m \leq i,j < n \wedge (i,j) \in grid(s,t) \wedge k \leq l < k+m \wedge$
    $X_{il} \neq 0 \wedge X_{lj} \neq 0$ do $X_{ij} := X_{ij} - X_{il}X_{lj}$
```

After the matrix update, the row and column nonzero counts must be adjusted because of the decreasing size of the reduced matrix and the creation of new nonzeros. It is sufficient to adjust the nonzero counts of those rows $i \geq k+m$ that have a nonzero entry X_{il} in a column l, $k \leq l < k+m$, and of those columns $j \geq k+m$ that have a nonzero entry X_{lj} in a row l, $k \leq l < k+m$. Nonzero counts of the other rows are not affected (for $i \geq k+m$) or are not needed any more (for $i < k+m$), and similarly for the column counts. The local data structure of column nonzero counts that is used to produce *SearchCols*(t) (see §2.2.1) must be adjusted accordingly.

The bulk of the computing work of the LU decomposition is formed by the main matrix-update loop; see the last statement of Algorithm 6. An efficient implementation strongly depends on the data structure used to represent the sparse matrix. This is the main subject of §3.

3. Local data structures.

In this section we analyse a number of candidate data structures for the local representation of a grid part of a sparse matrix. The candidates we consider are sparse data structures, which store only nonzeros. This leads to efficient use of computing time and memory. In our efficiency considerations, computing time is of primary importance, and memory use is of secondary importance. (Usually there is plenty of memory available on distributed-memory parallel computers.)

3.1. Description of data structures.

Grid parts of sparse matrices are sparse matrices themselves. Therefore, the data structure for a grid part of a sparse matrix can be chosen from the large variety of data structures that are used in sequential sparse matrix algorithms. Here, we consider several simple data structures, and describe them by Pascal-like type definitions. (A formal treatment of useful data structures is given in [30].) All data structures enable easy access to rows as well as to columns, which is required for sparse LU decomposition.

We examine data structures for the $\hat{n} \times \hat{n}$ sparse matrix $\hat{A}$ which represents the local grid part of the matrix A. Here, $\hat{n} = n/Q$. To simplify the analysis, we assume that $n \bmod Q = 0$. Local variables and indices are hatted, to distinguish them from global ones. For processor (s,t), the relation between $\hat{A}$ and A is given by

$$\hat{A}_{\hat{\imath}\hat{\jmath}} = A_{\hat{\imath}Q+s,\hat{\jmath}Q+t} \quad \text{for all } \hat{\imath},\hat{\jmath},\ 0 \le \hat{\imath},\hat{\jmath} < \hat{n}. \tag{7}$$

The data structures are:

1. *Gustavson's data structure* [20], *unordered.* The nonzeros of $\hat{A}$ are represented by array *entries.* The maximum number of nonzeros that can be stored is *bnd.* For each column, the nonzeros are stored in a contiguous block of *entries*, in arbitrary order. Between the blocks there can be empty space. For column cj of $\hat{A}$, *colfirst*[cj] gives the start of the block, and *lencol*[cj] its length. For each nonzero of column cj, both its row index $\hat{\imath}$ and its numerical value a are stored. In addition, there is a row data structure which is similar to the column data structure, except that numerical values of nonzeros are not stored.

```
type matrix = record
                rowfirst, colfirst : array [0..$\hat{n}-1$] of 0..$bnd-1$;
                lenrow, lencol     : array [0..$\hat{n}-1$] of 0..$\hat{n}$;
                $\hat{\jmath}$                  : array [0..$bnd-1$] of 0..$\hat{n}-1$;
                entries            : array [0..$bnd-1$] of entry
              end;
     entry  = record
                $\hat{\imath}$ : 0..$\hat{n}-1$;
                a : real
              end;
```

2. *Gustavson's data structure, ordered.* This data structure is the same as data structure 1, except that the nonzeros within each row and column are ordered by increasing index.

3. *Two-dimensional linked-list structure, unordered.* This dynamic data structure links the nonzeros of a row or column into a list, in arbitrary order. The headers of the $\hat{n}$ linked lists that correspond to the rows (columns) are represented by array *rows* (*cols*). Each entry has a pointer in its *next*$\hat{\jmath}$-field to another nonzero in the same row, and a pointer in its *next*$\hat{\imath}$-field to another nonzero in the same column.

```
type matrix       = record
                      rows, cols : array [0..$\hat{n}-1$] of entrypointer
                    end;
     entrypointer = ↑ entry;
     entry        = record
                      $\hat{\imath},\hat{\jmath}$          : 0..$\hat{n}-1$;
                      a            : real;
                      next$\hat{\imath}$, next$\hat{\jmath}$ : entrypointer
                    end;
```

4. *Two-dimensional linked-list structure, ordered.* This is the same data structure as the previous one, except that the nonzeros within each row and column list are ordered by increasing index. This data structure is similar to the orthogonal

linked list structure of Knuth [23]. The related Curtis–Reid data structure [7] can be obtained by leaving out the $\hat{\imath}$- and $\hat{\jmath}$-index of each entry and labelling the end of each row (column) list with the corresponding row (column) index. The Curtis–Reid data structure saves memory at the expense of an increase in computing time. We do not consider this data structure here, because memory use is not our primary concern.

5. *Two-dimensional doubly linked-list structure, unordered.* The idea of data structure 3 is carried further by representing the nonzeros of a row or column in a doubly linked list. This extension facilitates the deletion of an entry. We obtain data structure 5 by modifying the definition of *entry* in data structure 3:

```
type entry = record
        $\hat{\imath}, \hat{\jmath}$                            : $0..\hat{n}-1$;
        $a$                               : real;
        $prev\hat{\imath}, prev\hat{\jmath}, next\hat{\imath}, next\hat{\jmath}$ : entrypointer
    end;
```

6. *Two-dimensional doubly linked-list structure, ordered.* This is the same data structure as the previous one, except that the nonzeros within each row and column list are ordered by increasing index.

3.2. Computing time. We analyse the time complexity of a number of important computations for all data structures. Since the matrix is distributed over the processors, these computations are also distributed. Generally, a distributed computation includes some computation on the local matrix part by each processor, and some communication between the processors. Communication of rows and columns requires retrieval of these rows and columns from the matrix, followed by send and receive operations. Since the retrieval is similar to a local computation that is discussed below (multiple-row assignment) and since the send and receive operations are independent of the data structure, it is sufficient only to consider local computations. We examine the following local computations:

- *Multiple-row assignment* (see Algorithm 5).

```
for all $r : 0 \le r < ndest \wedge Dest(r) \bmod Q = s$ do
    for all $j : 0 \le j < n \wedge j \bmod Q = t \wedge (X_{Src(r),j} \ne 0 \vee X_{Dest(r),j} \ne 0)$
      do $X_{Dest(r),j} := X_{Src(r),j}$
```

- *Multiple-column division* (see Algorithm 6).

```
for all $i, j : k+m \le i < n \wedge k \le j < k+m \wedge (i,j) \in grid(s,t) \wedge X_{ij} \ne 0$
    do $X_{ij} := X_{ij}/X_{jj}$
```

- *Rank-m update* (see Algorithm 6).

```
for all $i, j, l : k+m \le i, j < n \wedge (i,j) \in grid(s,t) \wedge k \le l < k+m \wedge$
    $X_{il} \ne 0 \wedge X_{lj} \ne 0$ do $X_{ij} := X_{ij} - X_{il}X_{lj}$
```

Each of these local computations is performed once in every step of parallel sparse LU decomposition. Other local matrix computations that are performed within a single step, such as multiple-column assignment or multiple-column pivot search, very much resemble the above computations for all our data structures.

We evaluate the performance of the data structures by counting the number of operations in a local (sequential) computation by processor (s,t). We present a worst-case analysis, which gives upper bounds $\mathcal{O}(x)$ on the number of operations. The operation counts are expressed in global constants m and Q and in local constants $\hat{c}$ and $\hat{m}$; the maximum number of nonzeros in a row or column part is denoted by $\hat{c}$, and

TABLE 1
Operation counts $\mathcal{O}(x)$ for computations on local data structures.

Data structure	Multiple-row assignment	Multiple-column division	Rank-m update
1. Gustavson unordered	$\hat{c}^2\hat{m}$	$\hat{c}m/Q$	$\hat{c}^2m$
2. Gustavson ordered	$\hat{c}^2\hat{m}$	$\hat{c}m/Q$	$\hat{c}^2m(1+\log m)$
3. 2D list unordered	$\hat{c}^2\hat{m}$	$\hat{c}m/Q$	$\hat{c}^2m$
4. 2D list ordered	$\hat{c}^2\hat{m}$	$\hat{c}m/Q$	$\hat{c}^2m(1+\log m)$
5. 2D double list unordered	$\hat{c}\hat{m}$	$\hat{c}m/Q$	$\hat{c}^2m$
6. 2D double list ordered	$\hat{c}^2\hat{m}$	$\hat{c}m/Q$	$\hat{c}^2m(1+\log m)$

the maximum number of row parts $Dest(r)$, $0 \le r < ndest$, with $Dest(r) \bmod Q = s$, is denoted by $\hat{m}$. (It can be shown that $\hat{m} \le m + \lceil m/Q \rceil$, although it is more likely that $\hat{m} \approx 2m/Q$.) The actual performance in the parallel computation will deteriorate due to load imbalance, but in a manner which is independent of the data structure chosen. Therefore, our comparison of data structures remains valid for the parallel case.

Table 1 displays the orders of the operation counts for the different data structures. In the following, we shall briefly explain these results. The multiple-row assignment can be done by deleting the nonzeros of the old rows and inserting the nonzeros of the new rows. This involves at most $2\hat{m}$ row parts of at most $\hat{c}$ nonzeros, and hence the total number of nonzeros involved is $\mathcal{O}(\hat{c}\hat{m})$. The number of operations equals the number of nonzeros for the unordered data structure 5, since matrix elements can be assigned new values in $\mathcal{O}(1)$ operations: old elements are deleted by making use of the double links; new elements are inserted at the headers of the row and column lists. For the corresponding ordered data structure 6, insertion of a new element into a row takes $\mathcal{O}(\hat{c})$ operations, since its column predecessor must be found. This gives a total of $\mathcal{O}(\hat{c}^2\hat{m})$ operations. Similar considerations hold for the other data structures.

The multiple-column division takes $\mathcal{O}(\hat{c}m/Q)$ operations, since at most $\hat{c}$ nonzeros in at most $\lceil m/Q \rceil$ column parts are modified. This is the operation count for all the data structures, because they all allow column-wise access to the numerical values of the nonzeros.

For data structures 3–6, the rank-m update is a sequence of updates of target row parts. Each target row part i is updated by subtracting from it m update row parts l, each multiplied by a scalar X_{il}. This is done by: (i) scattering the nonzeros of the target row part into an array of length $\hat{n}$ that is known to be zero; (ii) scanning the update row parts and multiplying them by a scalar, while accumulating numerical values in the array and building a list of new nonzeros to be created; and (iii) updating the matrix data structure by adjusting numerical values and inserting new nonzeros, while resetting the array to zero. This is the second approach of [12, §2.4]. (In certain cases, such as updates by one row, it may be cheaper to scatter update rows into the array, instead of the target row. This is the first approach of [12, §2.4].) For data structures 1 and 2, the rank-m update is performed by columns, because of the column-wise access to the numerical values of the matrix.

For data structures 3–6, there are m column parts that contain nonzero multipliers X_{il}. Since each column part has at most $\hat{c}$ nonzeros, the total number of nonzero multipliers and hence of update row parts to be scanned is at most $\hat{c}m$. Since each

TABLE 2
Minimum memory requirement of data structures.

Data structure	Memory requirement per processor
1, 2. Gustavson	$4n/Q + 3cn/Q^2$
3, 4. 2D list	$2n/Q + 5cn/Q^2$
5, 6. 2D double list	$2n/Q + 7cn/Q^2$

update row part has at most $\hat{c}$ nonzeros, the total number of operations of (ii) is $\mathcal{O}(\hat{c}^2m)$. For data structures 1 and 2, the same upper bound holds.

In the unordered case, the total complexity of the rank-m update is $\mathcal{O}(\hat{c}^2m)$, because in (iii) new elements can be inserted in $\mathcal{O}(1)$ operations. This can be done at the end of row and column blocks for data structure 1, and at the headers of row and column lists for data structures 3 and 5.

In the ordered case, the nonzeros of the updated target row part must be obtained in order of increasing column index. This can be done by a symbolic merge-sort. In the worst case, m updates of a single target row part may cause it to grow from $\hat{c}$ to $\hat{c}(m+1)$ nonzeros. The corresponding merge-sort takes $\mathcal{O}(\hat{c}m \log m)$ operations. (All logarithms in this paper have base two.) In the worst case, this growth occurs for $\hat{c}$ target rows, so that the total number of operations in the merge-sorts is $\mathcal{O}(\hat{c}^2m \log m)$ and the total number of operations is $\mathcal{O}(\hat{c}^2m(1+\log m))$. (On average the differences between the data structures will be less pronounced than the last column of Table 1 suggests, because the growth rate of target rows may be less than the worst-case rate that has been assumed.)

3.3. Memory requirements. Table 2 shows the *minimum memory requirement* per processor for each data structure, i.e., the memory needed when the nonzeros are evenly distributed over the processors. We assume that the amount of memory per integer, real, and pointer is equal to 1. Memory requirements are expressed in terms of global constants n, c, and Q. The number of rows that appear in a grid part of the matrix is n/Q. The number of nonzeros in a row is assumed to be c, and the number of nonzeros in a row part is assumed to be c/Q.

All data structures require $\Omega(n/Q+cn/Q^2)$ memory. In the case of data structures 3–6, the memory needed for the n row headers and the n column headers scales with Q as $\Theta(Q^{-1})$, whereas the memory needed for the cn nonzeros scales as $\Omega(Q^{-2})$. This implies that row and column headers take up more memory in parallel computations than in sequential computations. Note that the grid distribution is still better in this respect than a row or column distribution, since in the latter case each processor would need $\Omega(n)$ memory, irrespective of Q. The scaling behaviour of data structures 1 and 2 is similar to that of the other data structures.

3.4. Discussion. Table 1 indicates that the unordered two-dimensional doubly linked-list structure 5 is superior: it outperforms all other data structures in the multiple-row assignment, and it is one of the three optimal data structures for the rank-m update. This conclusion is specific to *parallel* LU decomposition: in the sequential case explicit row permutations and hence row assignments often do not occur because permuting is done implicitly; furthermore, in the sequential case it is common to use rank-1 updates, and for $m = 1$ and $Q = 1$ the rank-m update takes $\mathcal{O}(c^2)$ operations for all data structures. Therefore, all data structures perform equally well sequentially.

A series of multiple-rank updates based on compatible pivot sets may be of benefit *even in the sequential case.* This gain is derived from updating an initial, small data structure in one large step, instead of updating a growing data structure in many small steps, each time accessing a larger and larger structure, with row length increasing from c to $c(m+1)$ in the worst case. For an unordered data structure, the sequential rank-m update has an operation count of $\mathcal{O}(c^2m)$, whereas a sequence of m rank-1 updates has a count of $\mathcal{O}(c^2m^2)$. Note that in this comparison, c is fixed as the number of nonzeros of a row *at the start* of the rank-m update or sequence of m rank-1 updates.

Table 1 shows that in all cases the time of the rank-m update dominates the time of the other computations. It also dominates the communication time: the number of communications in one step is of the order $\mathcal{O}(\hat{c}m + Q)$, since $\mathcal{O}(m)$ row parts of length $\hat{c}$ have to be communicated in each processor column of length Q, and similarly for column parts. For data structure 5, the communication count is about a factor $\hat{c}$ smaller than the operation count of the rank-m update.

The relative merit of data structures depends not only on operation counts, but also on other, sometimes machine-dependent, parameters. Linked-list data structures perform better on fast scalar processors, and hence on RISC-like architectures such as transputers. In contrast, the Gustavson data structure may perform better on vector processors.

We have chosen the conventional ordered two-dimensional linked-list structure 4 as the data structure for parallel sparse LU decomposition. Together with the grid distribution, data structure 4 has become the standard of all the parallel sparse linear algebra programs of PARPACK, a package developed at Koninklijke/Shell-Laboratorium, Amsterdam. This package includes among others LU decomposition, Cholesky factorisation, and triangular system solution. Our choice was made at a time when we did not recognise the importance of multiple-row assignments and rank-m updates in the context of parallel LU decomposition. Still, our data structure is conceptually simple, and it allows the development of programs with a reasonable amount of effort, so that it serves as an appropriate research vehicle. Also, it performs efficiently for a wide range of other algorithms, so that it is suitable as a compromise standard. As a suggestion for future research, we strongly encourage experiments with data structure 5, which according to our analysis is the most efficient data structure.

4. Experimental results.

The algorithm has been implemented in the parallel programming language occam 2 (for an introduction, see [4]) and experimental results have been obtained on a Parsytec SuperCluster FT-400 parallel computer. This machine consists of a square mesh of 400 INMOS T800-20 transputers, each with a 2 Mbyte memory. According to our measurements, a transputer performs a 64-bit floating point operation (flop), such as addition or multiplication, in $t_{\mathrm{flop}} = 1.9\,\mu\mathrm{s}$, and a 32-bit integer operation in $t_{\mathrm{iop}} = 1.8\,\mu\mathrm{s}$. A transputer sends a 64-bit real number to a neighbouring transputer in the mesh in $t_{\mathrm{comm,real}} = 8.5\,\mu\mathrm{s}$ and it sends a 32-bit integer in $t_{\mathrm{comm,int}} = 6.6\,\mu\mathrm{s}$. In the experiments, all real numbers have a length of 64 bits, and all integers have a length of 32 bits. The machine accuracy for 64-bit floating point operations is 2.2×10^{-16}. All experimental times were measured by an internal timer calibrated with a wall clock.

TABLE 3
Test set of sparse matrices.

Matrix	Origin	n	$nz(A)$	$c(A)$	$c(L\backslash U)$
IMPCOL B	Chemical engineering	59	312	5.3	7.4
WEST0067	Chemical engineering	67	294	4.4	8.7
FS 541 1	Atmospheric pollution	541	4285	7.9	30.7
STEAM2	Oil reservoir simulation	600	13760	22.9	89.0
SHL 400	Linear programming	663	1712	2.6	2.6
BP 1600	Linear programming	822	4841	5.9	8.8
JPWH 991	Electronic circuit simulation	991	6027	6.1	66.3
SHERMAN1	Oil reservoir simulation	1000	3750	3.8	26.6
SHERMAN2	Oil reservoir simulation	1080	23094	21.4	326.2
LNS 3937	Compressible fluid flow	3937	25407	6.5	107.0
GEMAT11	Optimal power flow	4929	33185	6.7	10.8

4.1. Test set of Harwell–Boeing matrices. To investigate the properties of our algorithm, we performed numerical experiments on eleven real unsymmetric assembled (RUA) matrices from the Harwell–Boeing sparse matrix collection [13]. In our test set we included matrices from diverse application fields, with widely varying sizes and nonzero densities. Fig. 1(a) shows the matrix IMPCOL B from the test set. Table 3 presents data on the test matrices: n is the matrix order; $nz(A)$ is the number of nonzeros of the matrix A before LU decomposition; $c(A)$ is the corresponding average number of nonzeros per row; $c(L\backslash U)$ is the average number of nonzeros per row of the matrix $L\backslash U$ that contains the L and U factors of A. These L and U factors were obtained by executing the parallel program on 400 processors.

4.2. Algorithm with standard parallel pivot search strategy. In this subsection we compare the parallel program running on $p = Q^2$ transputers, $1 \leq p \leq 400$, with a sequential program running on one transputer. The parallel program is an implementation of the algorithm of §2. The data structure is the ordered two-dimensional linked-list structure, i.e., data structure 4 of §3. Our standard pivot search procedure is a specific implementation of the parallel algorithm of §2.2 for $ncol = 1$. (Alternative pivot search procedures will be examined in §4.3.) Candidate pivot elements must satisfy (3) with $u = 0.1$, as recommended by Duff, Erisman, and Reid [12, Chap. 7]. Candidates are rejected on sparsity grounds if their Markowitz count is higher than four times the minimum Markowitz count of the candidates, as in the experiments of Davis and Yew [9].

The sequential program is a well-optimised version of the parallel program. It is obtained by simplifying the parallel program, removing all parallel overhead, and wherever possible exploiting the fact that $p = 1$. The parallel pivot search strategy is replaced by the common sequential strategy of searching three matrix columns for numerically acceptable pivot candidates and then choosing one candidate with the lowest Markowitz count.

The sequential and $p = 1$ times for matrices SHERMAN2 and LNS 3937 had to be obtained on a separate transputer with 16 Mbyte memory, because these problems do not fit into the 2 Mbyte memory of a transputer of the FT-400. (The maximum number of nonzeros per processor is about 70,000.) Incidentally, the separate transputer is about 1.13 times faster on sparse LU decomposition problems than a transputer

TABLE 4
Time (in s) of parallel sparse LU decomposition on p transputers.

Matrix	seq	$p=1$	4	9	16	25	49	100	400
IMPCOL B	0.34	0.38	0.20	0.15	0.13	0.12	0.12	0.11	0.13
WEST0067	0.47	0.53	0.27	0.20	0.18	0.18	0.15	0.13	0.16
FS 541 1	18.4	18.8	6.74	4.47	2.91	2.25	1.79	1.42	1.29
STEAM2	92.9	93.9	28.2	14.1	9.31	7.15	4.82	3.59	2.58
SHL 400	2.31	2.69	1.40	1.05	0.88	0.82	0.73	0.71	0.66
BP 1600	9.97	10.7	4.46	3.30	2.52	2.20	1.94	1.79	1.74
JPWH 991	121.	131.	34.8	20.4	13.7	10.3	7.33	5.50	3.97
SHERMAN1	36.6	41.0	13.0	7.00	6.08	4.33	3.09	2.75	2.37
SHERMAN2	1668.	1592.		196.	108.	87.8	46.2	32.7	15.6
LNS 3937	2119.	2111.		261.	168.	96.7	77.2	37.9	23.7
GEMAT11	75.6	84.3	29.3	18.2	14.1	11.4	9.19	7.73	6.23

of the FT-400. Times measured on the separate transputer were multiplied by 1.13, to obtain comparable table entries. The $p = 4$ times for matrices SHERMAN2 and LNS 3937 could not be obtained, because we did not have a four-transputer network available with sufficient memory.

Table 4 shows the time T_p of parallel LU decomposition on p processors and the time T_{seq} of sequential decomposition. The scaling behaviour of T_p with p can be explained qualitatively by the expression

$$T_p = \mathcal{O}\left(\frac{c^2 n}{p} + \frac{cn}{\sqrt{p}} + \frac{\sqrt{p}n}{m}\right), \tag{8}$$

where c is the average number of nonzeros per row during the LU decomposition and m is the average rank of a matrix update. The first term represents the computing time of n/m rank-m updates, each of time $\mathcal{O}(\hat{c}^2 m(1+\log m)) \approx \mathcal{O}(c^2 m/p)$; see Table 1. (In this rough approximation the $m \log m$ term is neglected.) The second term represents mainly the time needed to communicate row and column parts in n/m steps, each of time $\mathcal{O}(\hat{c}m) \approx \mathcal{O}(cm/\sqrt{p})$; see §3.4. The third term includes the startup time of communication pipelines in n/m steps, each of time $\mathcal{O}(\sqrt{p})$. A few smaller terms have been neglected in the derivation of (8). Usually the first term dominates for small p, the second term for intermediate p, and the third term for large p. The cross-over points between these ranges are problem-dependent.

The table shows that T_p is a monotonically decreasing function of p, with the exception of the increase that occurs in moving from $p = 100$ to $p = 400$ for the matrices IMPCOL B and WEST0067. This increase is hardly surprising, since at the start of the computation there are fewer nonzeros (312 or 294) than processors (400), so there is little to compute per processor and the total time is dominated by the third term of (8), which increases with p. In the other cases either the first or the second term dominates. For example, the gain by a factor of two in computing rate for SHERMAN2 from $p = 100$ to $p = 400$ is probably due to the dominant behaviour of the second term.

The maximum speedup $S_p = T_{\text{seq}}/T_p$ achieved is $S_{400} \approx 107$ for SHERMAN2. Tables 3 and 4 show that the speedup is correlated to $c(L\backslash U)$: speedups increase with increasing $c(L\backslash U)$.

TABLE 5
Number of steps of parallel sparse LU decomposition on p transputers.

Matrix	$p = 1$	4	9	16	25	49	100	400
	seq							
IMPCOL B	59	35	30	25	22	20	14	15
WEST0067	67	40	33	32	27	24	20	18
FS 541 1	541	318	238	212	179	154	127	104
STEAM2	600	404	337	306	257	228	220	181
SHL 400	663	332	223	171	139	105	83	45
BP 1600	822	561	465	399	347	314	268	190
JPWH 991	991	679	558	506	477	398	376	306
SHERMAN1	1000	616	430	405	327	274	250	200
SHERMAN2	1080		911	857	795	803	785	745
LNS 3937	3937		2252	2140	1830	1634	1318	1237
GEMAT11	4929	2589	1799	1389	1152	890	662	386

Table 4 shows that the difference in running time between the sequential program and the parallel program with $p = 1$ is small. The difference is due partly to parallel overhead and partly to differing pivot search strategies: the sequential program searches three columns per step, whereas the parallel program with $p = 1$ searches only $Q \cdot ncol = 1$ column. In most cases, the parallel algorithm is slower due to the parallel overhead and the lower quality (with respect to fill-in reduction) of the pivots. In two cases, SHERMAN2 and LNS 3937, the parallel algorithm is faster because these slowdown effects are more than offset by a faster search for pivots.

Table 5 shows the number of steps of the LU decomposition. This number is at least $n/\sqrt{p}$, because at most $\sqrt{p}$ compatible pivot elements are found in each step, due to our choice of $ncol = 1$. The near-ideal behaviour shown by matrices SHL 400 and GEMAT11 can be explained as follows. For a general $n \times n$ matrix with c nonzeros per row, the probability of an arbitrary element being zero is $1 - c/n$, so that the probability of two arbitrary pivot candidates being compatible is $(1 - c/n)^2$; cf. (4). This probability is even higher for pivot candidates that are chosen from the sparsest columns of the matrix, as in our pivot search strategy. Both SHL 400 and GEMAT11 have a very low nonzero density c/n, before and during LU decomposition. This implies that pivot candidates are usually compatible, so that most candidates become pivots and the number of steps is close to minimum. (Note that the number of steps depends on the ratio c/n and not on c alone.)

For all matrices, the number of steps initially decreases rapidly with increasing p, until a saturation point is reached. This point represents the situation where the algorithm proceeds through the sparse part of the matrix in a few steps of high rank m, and then handles the remaining dense part in steps of rank $m = 1$; see Fig. 1(d).

Table 6 displays the number of nonzeros and floating point operations, and the numerical error of the LU decomposition. The number of nonzeros fluctuates with p, without a clear trend, and without a clear advantage to either the sequential or the parallel program. From these results, we conclude that there is at most a limited penalty in terms of fill-in for relaxing the original Markowitz pivot search strategy.

The number of floating point operations is obtained by incrementing a counter for every flop performed, including the redundant flops that are introduced by parallelising the sequential program. The growth in flop count that can be seen for the very

TABLE 6

Number of nonzeros and floating point operations, and numerical error of parallel sparse LU decomposition on p transputers.

Matrix	Nonzeros $nz(L\backslash U)$			Floating point operations			Numerical error		
	seq	$p = 16$	$p = 400$	seq	$p = 16$	$p = 400$	seq	$p = 16$	$p = 400$
IMPCOL B	412	445	435	8.8e+2	1.9e+3	3.9e+3	3e−12	2e−12	3e−13
WEST0067	544	607	583	1.9e+3	4.0e+3	8.3e+3	7e−15	1e−14	3e−14
FS 541 1	13553	15056	16609	2.2e+5	3.1e+5	5.8e+5	6e−15	7e−15	3e−15
STEAM2	42869	45010	53395	1.7e+6	2.1e+6	4.1e+6	6e−9	6e−9	8e−9
SHL 400	1712	1712	1712	0	4.0e+1	2.2e+3	0	0	0
BP 1600	7424	7398	7229	2.6e+4	3.1e+4	7.2e+4	5e−12	4e−12	7e−12
JPWH 991	68587	69263	65707	6.9e+6	7.7e+6	7.1e+6	9e−12	3e−12	1e−11
SHERMAN1	27089	30989	26602	1.4e+6	1.8e+6	1.5e+6	1e−11	7e−12	2e−11
SHERMAN2	375316	348651	352301	8.0e+7	6.9e+7	8.4e+7	6e−7	8e−6	4e−6
LNS 3937	474800	448929	421090	8.3e+7	7.4e+7	7.4e+7	2e−4	1e−4	2e−3
GEMAT11	53358	54086	53093	4.8e+5	5.8e+5	9.3e+5	2e−10	4e−10	3e−11

sparse matrix SHL 400 is entirely due to such redundant computations. For denser matrices the redundancy is negligible. In most cases, increasing p leads to a limited growth in flop count.

The numerical error is defined as the maximum absolute error in the numerical solution x of a system $Ax = b$, with b chosen such that all components of the exact solution are one. The system has not been scaled. It has been solved in stages 1–5; see §1. The error shows variations of one order of magnitude, without any distinguishable trend. We attribute this behaviour to coincidence (e.g., caused by arbitrary tie breaking in the pivot search) and not to the qualities of any particular pivot strategy. In most cases the accuracy of our program is comparable to that of the programs MA28 [11] and D2 [8], [9], with the notable exception of the cases STEAM2, SHERMAN2, and LNS 3937, for which our program is less accurate (cf. [8, Table 4.3]). The accuracy can be improved by appropriately increasing u, decreasing a, and increasing *ncol*. This has been confirmed by experiments, except for LNS 3937, which could not be solved with a higher accuracy than 10^{-4}, whatever the choice of parameters. (For this matrix, however, accuracy can be improved significantly by scaling.)

4.3. Algorithm with alternative pivot search strategies. To investigate the influence of the chosen pivot strategy, we replaced the pivot search procedure by alternative search procedures. The other parts of the program remained unchanged. The first alternative is to search one column per processor column, and then choose *one* pivot with the lowest Markowitz count from the Q pivot candidates. This leads to a rank-1 update in each step, which is the usual method in sequential algorithms. The second alternative is a general implementation of the pivot search algorithm of §2.2 for arbitrary *ncol*. We present results for the case $ncol = 3$. Fig. 1 illustrates the LU decomposition of IMPCOL B for $ncol = 5$ and $Q = 2$.

Figure 2 shows the time of parallel sparse LU decomposition for the standard pivot search strategy ($ncol = 1$) and the two alternative strategies (rank-1, $ncol = 3$). We observe that the standard strategy is clearly superior to the rank-1 strategy, in particular for matrices with low density c/n, such as GEMAT11. This is due to the exploitation of sparsity-based parallelism, which is not used in the rank-1 strategy. The standard strategy is also better than the $ncol = 3$ strategy. We attribute this mainly to the simplicity of the pivot search with *ncol* = 1, which requires less communication than the search with *ncol* = 3. The potential gain in parallelism caused by increasing the number of pivots (which is at most $m_{\max} = \sqrt{p} \cdot ncol$), and hence increasing the rank of the updates, does not compensate for the losses incurred during the more expensive pivot search.

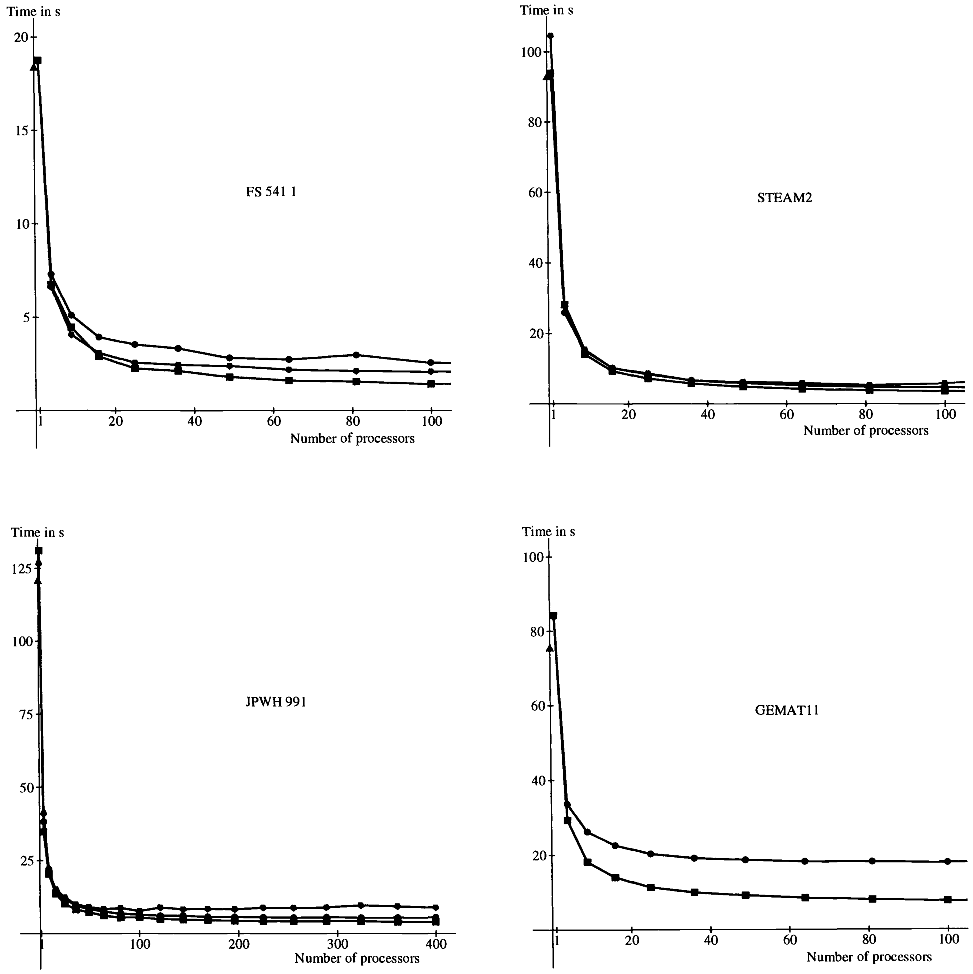


FIG. 2. *Time of parallel sparse LU decomposition for three pivot search strategies. The test matrices are* (a) FS 541 1; (b) STEAM2; (c) JPWH991; *and* (d) GEMAT11. *The squares denote the standard pivot search strategy* (*with ncol* = 1); *the pentagons the strategy with ncol* = 3; *and the circles the rank*-1 *strategy. The time of the sequential program is shown by a triangle. For* GEMAT11 *the curves of ncol* = 1 *and ncol* = 3 *nearly coincide, and only the first curve is shown.*

A possible application of the pivot search strategy with high *ncol* is when several matrices have to be decomposed that have an identical sparsity pattern but slightly different numerical values. In that case, the permutations π and ρ of (2) are determined during the decomposition of the first matrix. The remaining matrices are ordered according to π and ρ and then decomposed without pivot searching or permuting. This saves the parallel pivot search and the parallel permutations which may consume up to 80 per cent of the computing time. This procedure has been shown to lead to significant gains in the sequential case [12, Chap. 5], and also in the parallel case [28]. In our case, an initial investment in determining a high-quality ordering by using the alternative pivot strategy with high *ncol* (and low *a* and high *u*) would pay off handsomely in subsequent decompositions.

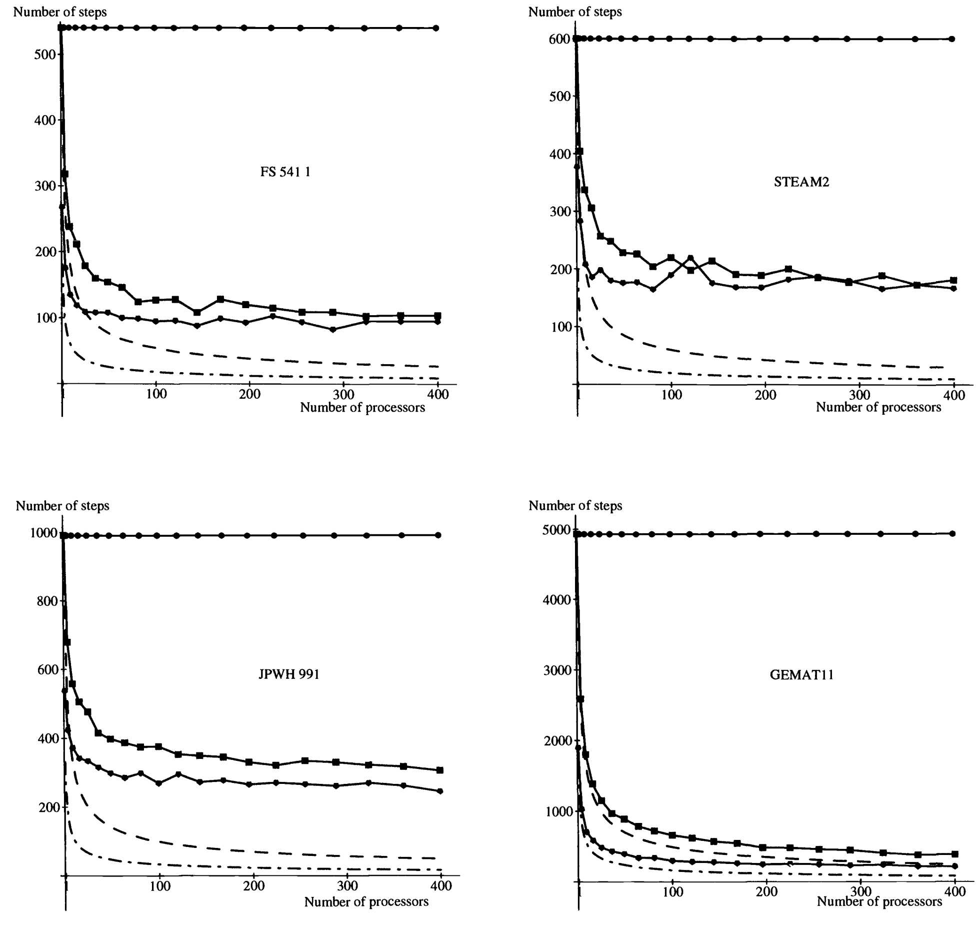


FIG. 3. *Number of steps of parallel sparse LU decomposition for three pivot search strategies. The test matrices and the marker symbols are the same as in Fig. 2. The dashed line denotes the minimum number of steps, $n/\sqrt{p}$, for the standard strategy; and the dashed-dotted line, the minimum number, $n/3\sqrt{p}$, for the ncol = 3 strategy.*

Figure 3 shows the number of steps of parallel sparse LU decomposition for the standard strategy and the two alternatives. The rank-1 strategy gives n steps in all cases. The *ncol* = 3 strategy gives a smaller number of steps than the standard strategy, in particular for a small number of processors. For a larger number of processors the difference between these two strategies is relatively small. Note that the curves of the matrix GEMAT11 are close to the ideal curves.

The results with respect to number of nonzeros, number of floating point operations, and numerical error of the alternative pivot strategies are similar to those of the standard strategy.

5. Conclusions. In this paper, we have presented a scalable parallel sparse LU decomposition algorithm which is based on the grid distribution and the ordered two-dimensional linked-list data structure. The algorithm scales reasonably well with the number of processors, and it achieves a speedup of up to 107 on 400 processors for large problems, i.e., problems with a large matrix order n and a high average number c of nonzeros per row. The potential of the algorithm is best exploited in the solution of large problems, such as the larger problems in the Harwell–Boeing collection.

Our general-purpose algorithm exploits both density-based and sparsity-based parallelism. In a way, these two kinds of parallelism supplement each other: matrices with high c have many potentially simultaneous operations in each rank-1 update; matrices with low nonzero density c/n have many potentially simultaneous rank-1 updates. Matrices with high c and low c/n benefit from both kinds of parallelism. Matrices with low c but high c/n (and hence small n) offer little hope for parallelism.

The timing results of our sparse LU decomposition program on 400 transputers show that a distributed-memory parallel computer can successfully compete in the field of sparse matrix computations with today's fastest uniprocessor supercomputers. As an example, the problem SHERMAN2 is solved in 15.6 s by our program running on the FT-400, and in 34.4 s by MA28 running on one processor of the CRAY YMP/832 [27]. (This is only an indication of relative speeds, as computing speeds are obviously problem-dependent: for SHERMAN1 the CRAY YMP/832 is four times faster than the FT-400.)

Future research may lead to significant improvement of the algorithm of this paper. First, the data structure can be changed into the unordered two-dimensional doubly linked-list structure 5, which theoretically has the lowest time complexity; see §3. Second, the distributed-memory parallel sparse algorithm can be combined with a parallel dense algorithm [3] which is invoked as soon as the nonzero density of the reduced matrix exceeds a certain value and sufficient memory is available to store the reduced matrix as a dense matrix. Such a switch from a sparse to a dense program is performed in the shared-memory parallel algorithm D2 [9] (when $c/n \geq 0.2$), and in the shared-memory parallel algorithms Y12M1, Y12M2, and Y12M3 [17] (when $c/n \geq 0.1$), with often large gains. This switch prevents, for instance, extensive searching for compatible pivots when only few compatible pivots exist due to the high density of the reduced matrix. Third, several pivot search strategies can be combined: searching for large pivot sets in the first few steps of the algorithm ($ncol > 1$), then searching for smaller sets ($ncol = 1$), and after that searching for single pivots, and finally switching to a dense algorithm. (A similar combination of strategies is proposed in [17].) We expect future hybrid algorithms with appropriate switch-over criteria to be considerably faster than the current algorithm.

Appendix. Postcondition and invariant. The aim of the algorithm for processor (s, t) is to establish the *postcondition* [10] $R[s, t]$, which is the following logical expression:

$$\begin{array}{llll} R[s,t]: & \forall_{i,j\,:\,i\leq j\,\wedge\,(i,j)\in grid(s,t)} & X_{ij} = U_{ij} = A_{\pi_i,\rho_j} - \displaystyle\sum_{h=0}^{i-1} L_{ih}U_{hj} \\ \wedge & \forall_{i,j\,:\,i>j\,\wedge\,(i,j)\in grid(s,t)} & X_{ij} = L_{ij} = \left(A_{\pi_i,\rho_j} - \displaystyle\sum_{h=0}^{j-1} L_{ih}U_{hj} \right) / U_{jj}. \end{array}$$

At the start of step k of the algorithm, processor (s,t) guarantees the truth of the following *invariant* [10]:

$$\begin{array}{lll} P[s,t]: & \forall_{i,j\,:\,i,j\geq k\,\wedge\,(i,j)\in grid(s,t)} & X_{ij} = A_{\pi_i,\rho_j} - \displaystyle\sum_{h=0}^{k-1} X_{ih}X_{hj} \\ \wedge & \forall_{i,j\,:\,i<k\,\wedge\,i\leq j\,\wedge\,(i,j)\in grid(s,t)} & X_{ij} = A_{\pi_i,\rho_j} - \displaystyle\sum_{h=0}^{i-1} X_{ih}X_{hj} \\ \wedge & \forall_{i,j\,:\,j<k\,\wedge\,i>j\,\wedge\,(i,j)\in grid(s,t)} & X_{ij} = \left(A_{\pi_i,\rho_j} - \displaystyle\sum_{h=0}^{j-1} X_{ih}X_{hj} \right) / X_{jj} \\ \wedge & \forall_{i\,:\,i\geq k\,\wedge\,i \bmod Q = s} & R_i = |\{j : k \leq j < n \wedge X_{ij} \neq 0\}| \\ \wedge & \forall_{j\,:\,j\geq k\,\wedge\,j \bmod Q = t} & C_j = |\{i : k \leq i < n \wedge X_{ij} \neq 0\}|. \end{array}$$

The first three subexpressions of the invariant state which partial sums have been accumulated so far. The last two subexpressions describe the nonzero counts. All invariants $P[s,t]$ hold trivially for $k = 0$, after the initialisation $X = A$ and $\pi = \rho =$ id and the appropriate initialisation of the nonzero count vectors R and C. At the end of the computation, for $k = n$, all components have their final values, so that the postcondition $R[s,t]$ is established. The algorithms works towards $R[s,t]$ by repeatedly incrementing k from 0 to n, while keeping the invariants $P[s,t]$ valid.

Acknowledgments. We are grateful to Daniël Loyens and John Somers for many stimulating discussions, and also for providing us with the Interactive Linear Algebra System (ILIAS), which greatly facilitated our numerical experiments. We thank Rob Miesen for help in plotting the figures.

REFERENCES

[1] G. Alaghband, *Parallel pivoting combined with parallel reduction and fill-in control*, Parallel Comput., 11 (1989), pp. 201–221.

[2] R. H. Bisseling and J. G. G. van de Vorst, *Parallel triangular system solving on a mesh network of transputers*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 787–799.

[3] ———, *Parallel* LU *decomposition on a transputer network*, in Lecture Notes in Computer Science 384, Springer-Verlag, New York, 1989, pp. 61–77.

[4] A. Burns, *Programming in occam* 2, Addison-Wesley, Wokingham, U.K., 1988.

[5] D. A. Calahan, *Parallel solution of sparse simultaneous linear equations*, in Proc. 11th Annual Allerton Conf. on Circuits and System Theory, 1973, pp. 729–735.

[6] E. Chu and A. George, *Gaussian elimination with partial pivoting and load balancing on a multiprocessor*, Parallel Comput., 5 (1987), pp. 65–74.

[7] A. R. Curtis and J. K. Reid, *The solution of large sparse unsymmetric systems of linear equations*, J. Inst. Math. Appl., 8 (1971), pp. 344–353.

[8] T. A. Davis, *A parallel algorithm for sparse unsymmetric* LU *factorization*, Ph. D. thesis, Tech. Rep. 907, Center for Supercomputing Research and Development, Univ. of Illinois, Urbana, IL, Sept. 1989.

[9] T. A. Davis and P.-C. Yew, *A nondeterministic parallel algorithm for general unsymmetric sparse* LU *factorization*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 383–402.

[10] E. W. Dijkstra, *A Discipline of Programming*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[11] I. S. Duff and J. K. Reid, *Some design features of a sparse matrix code*, ACM Trans. Math. Software, 5 (1979), pp. 18–35.

[12] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*, Oxford University Press, Oxford, U.K., 1986.

[13] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[14] A. M. Erisman, R. G. Grimes, J. G. Lewis, W. G. Poole, Jr., and H. D. Simon, *Evaluation of orderings for unsymmetric sparse matrices*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 600–624.

[15] G. C. Fox, M. A. Johnson, G. A. Lyzenga, S. W. Otto, J. K. Salmon, and D. W. Walker, *Solving Problems on Concurrent Processors*, Vol. 1, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[16] K. A. Gallivan, B. A. Marsolf, and H. A. G. Wijshoff, MCSPARSE: *A parallel sparse unsymmetric linear system solver*, Tech. Rep. 1142, Center for Supercomputing Research and Development, Univ. of Illinois, Urbana, IL, Aug. 1991.

[17] K. Gallivan, A. Sameh, and Z. Zlatev, *Parallel direct method codes for general sparse matrices*, Tech. Rep. 1143, Center for Supercomputing Research and Development, Univ. of Illinois, Urbana, IL, Oct. 1991.

[18] G. A. Geist and C. H. Romine, LU *factorization algorithms on distributed-memory multiprocessor architectures*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 639–649.

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

[20] F. G. Gustavson, *Some basic techniques for solving sparse systems of linear equations*, in Sparse Matrices and Their Applications, D. J. Rose and R. A. Willoughby, eds., Plenum Press, New York, 1972, pp. 41–52.

[21] M. T. Heath, E. Ng, and B. W. Peyton, *Parallel algorithms for sparse linear systems*, SIAM Rev., 33 (1991), pp. 420–460.

[22] S. L. Johnsson, *Communication efficient basic linear algebra computations on hypercube architectures*, J. Parallel Distrib. Comput., 4 (1987), pp. 133–172.

[23] D. E. Knuth, *The Art of Computer Programming*, Vol. 1, 2nd ed., Addison-Wesley, Reading, MA, 1973.

[24] H. M. Markowitz, *The elimination form of the inverse and its application to linear programming*, Management Sci., 3 (1957), pp. 255–269.

[25] D. M. Nicol and J. H. Saltz, *An analysis of scatter decomposition*, IEEE Trans. Comput., 39 (1990), pp. 1337–1345.

[26] P. Sadayappan and S. K. Rao, *Communication reduction for distributed sparse matrix factorization on a processor mesh*, in Proc. Supercomputing '89, ACM Press, New York, 1989, pp. 371–379.

[27] M. K. Seager, *A* SLAP *for the masses*, in Parallel Supercomputing: Methods, Algorithms and Applications, G. F. Carey, ed., John Wiley, Chichester, U.K., 1989, pp. 135–155.

[28] A. Skjellum, *Concurrent dynamic simulation: Multicomputer algorithms research applied to ordinary differential-algebraic process systems in chemical engineering*, Ph. D. thesis, California Institute of Technology, Pasadena, CA, May 1990.

[29] D. Smart and J. White, *Reducing the parallel solution time of sparse circuit matrices using reordered Gaussian elimination and relaxation*, in Proc. IEEE Internat. Symp. Circuits and Systems, 1988, pp. 627–630.

[30] A. F. van der Stappen, *Distributed data structures for sparse linear algebra*, Master's thesis, Eindhoven Univ. of Technology, the Netherlands, June 1988.

[31] E. F. Van de Velde, *Experiments with multicomputer* LU*-decomposition*, Concurrency: Practice and Experience, 2 (1990), pp. 1–26.

[32] J. G. G. van de Vorst, *The formal development of a parallel program performing* LU*-decomposition*, Acta Inform., 26 (1988), pp. 1–17.

[33] Z. Zlatev, *On some pivotal strategies in Gaussian elimination by sparse technique*, SIAM J. Numer. Anal., 17 (1980), pp. 18–30.

[34] ———, *Computational Methods for General Sparse Matrices*, Mathematics and Its Applications 65, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1991.

[35] Z. Zlatev, J. Wasniewski, and K. Schaumburg, Y12M—*Solution of Large and Sparse Systems of Linear Algebraic Equations*, Lecture Notes in Computer Science 121, Springer-Verlag, Berlin, 1981.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 3, pp. 880–901, July 1993

018

COMPARTMENTAL MODELING AND SECOND-MOMENT ANALYSIS OF STATE SPACE SYSTEMS*

DENNIS S. BERNSTEIN† AND DAVID C. HYLAND‡

Abstract. Compartmental models involve nonnegative state variables that exchange mass, energy, or other quantities in accordance with conservation laws. Such models are widespread in biology and economics. In this paper a connection is made between arbitrary (not necessarily nonnegative) state space systems and compartmental models. Specifically, for an arbitrary state space model with additive white noise, the nonnegative-definite second-moment matrix is characterized by a Lyapunov differential equation. Kronecker and Hadamard (Schur) matrix algebra is then used to derive an equation that characterizes the dynamics of the diagonal elements of the second-moment matrix. Since these diagonal elements are nonnegative, they can be viewed, in certain cases, as the state variables of a compartmental model. This paper examines weak coupling conditions under which the steady-state values of the diagonal elements actually satisfy a steady-state compartmental model.

Key words. stochastic models, power flow, nonnegative matrices

AMS subject classifications. 15, 15A45

1. Introduction.

Analysis and design methodologies based upon worst-case behavior can be unduly pessimistic for applications in which system behavior includes highly improbable events. It is thus our goal to undertake a probabilistic approach to account for (or, more aptly, to *ignore*) unlikely behavior to achieve higher performance and more realistic predictions. Accordingly, we consider an H_2/white noise (as opposed to an H_∞/L_2) system and signal model as a starting point. Now, however, we seek system models that ignore detailed *microscopic* modeling data while focussing on the most likely *macroscopic* phenomena. Our paradigm is heat flow in which molecular motion is highly uncertain, whereas energy flows, with virtual certainty, from hot objects to cold objects.

Most probable motion in dynamical systems is the traditional province of statistical mechanics, which normally deals with very large (say, 10^{23}) interacting components. Our challenge in the field of modeling for robust control is to develop a useful theory of "statistical mechanics of moderate-sized systems." Such a theory does not currently exist due to the emphasis by physicists on large stochastic systems as well as the emphasis by engineers, dynamicists, and control theorists on relatively small deterministic systems. It is our view that a "middle ground theory" is needed to fill the gap between these worlds. The benefits of such a theory include the means to overcome the inherent limitations of worst-case design. The present paper is directed toward this goal.

To begin we shall focus on dynamical systems that involve subsystems or states whose values are nonnegative quantities [1]–[13]. Dynamical models of such systems are based upon the physics of the processes by which various quantities are exchanged by the coupled subsystems. In addition, conservation laws are used to account for the possibly macroscopic transfer (or flow) of such quantities among subsystems. Models for this class of systems are known as *compartmental* models.

The range of application of compartmental models is quite large. Their usage is widespread in biology and ecology [10], [12], while closely related ideas appear in economics [6, Chap. 9]. Our interest in compartmental models arises from electrical and mechanical engineering applications. Thus far there has been little direct connection between these engineering disciplines and compartmental modeling since classical

* Received by the editors April 22, 1991; accepted for publication (in revised form) October 14, 1992.

† Department of Aerospace Engineering, University of Michigan, Ann Arbor, Michigan 48109-2140 (dsbaero@caen.engin.umich.edu).

‡ Harris Corporation, Mail Stop 22/4847, Melbourne, Florida 32902.

(R, L, C) circuit models and (M, D, K) structural models are not cast in terms of inherently nonnegative quantities and do not explicitly invoke conservation laws.

The goal of this paper is to demonstrate a direct connection between arbitrary linear dynamical systems in state space form and compartmental models. The key to this connection is the recognition that even for arbitrary systems that do not explicitly involve nonnegative quantities (such as the (R, L, C) and (M, D, K) models mentioned above), it is possible to identify nonnegative quantities that do behave like compartmental models with conservation laws. In certain specific cases such connections have already been demonstrated, albeit, usually without recognition of compartmental concepts. Examples include energy flow and power transfer in random media [14]–[17], dissipative circuits [18]–[27], mechanical systems [28]–[30], coupled structures [31]–[46], and networks of queues [47] and [48]. A compartmental-like description of coupled structures is given in [35].

In each of the above applications, the key to formulating the system dynamics in compartmental form is to characterize nonnegative quantities that arise from the underlying physical phenomena. The reason that such models have not been more widely used is that physical principles such as Kirchhoff's laws and Newton's law are not usually formulated in terms of subsystem interaction and energy transfer. However, once the underlying laws of physics have been formulated as a dynamical system, it is often possible to reformulate these dynamics in terms of energy transfer. There are at least three mathematical formulations that may give rise to compartmental models:

(i) root mean square (rms) averaging of system states over time within a deterministic formulation,

(ii) averaging system states over the statistics of stochastic disturbances, and

(iii) averaging system states over the statistics of uncertain parameters.

The nonnegative quantities that arise from these formulations are then simply the mean-square averages of the original not-necessarily nonnegative states. In this paper we consider (ii), while approach (iii) underlies much of Statistical Energy Analysis [31]–[42] and has been explored in [49] and [50]. The averaging techniques developed in [51]–[53] are also related to (iii). The results of this paper may also be applicable to large-scale systems problems [54]–[59]. Such connections remain to be explored.

The goal of this paper is to establish some basic mathematical results that demonstrate how compartmental models arise from a second-moment analysis of state space systems. Physical interpretation of the derived compartmental models will not concern us here, while connections to circuit theory and dynamics will be explored elsewhere. Indeed, the above allusions to electrical and mechanical systems should be viewed as purely motivational. Within the paper we shall, however, use "energy" and "power" terminology as generic language to facilitate the discussion.

After introducing some global notation at the end of this section, we proceed in § 2 to summarize some basic properties of compartmental models. Using [6] as our principal reference, we show that compartmental models are confined to a nonnegative state space (Proposition 2.1) and then give necessary and sufficient conditions for the existence of a steady-state equilibrium energy distribution (Proposition 2.2). Sufficient conditions are also given (Corollary 2.2) under which the steady-state distribution is uniform. This phenomenon is known as "equipartition of energy" [34], [41], [50] and is also related to the notion of a "monotemperaturic" system [21]. We stress that although many of these results are well known [3], [6], they are restated here in a concise and unified format that supports the development in later sections.

Specializing to the asymptotically stable case, we then consider the problem of determining the steady-state energy distribution in the limit of strong coupling, that is, the case in which the off-diagonal terms in the dynamics matrix become arbitrarily large (Proposition 2.4). As a special case of this result we state conditions under which energy equipartition occurs (Corollary 2.3).

In § 3 we shift gears and undertake an analysis of the nonnegative-definite[1] second-moment matrix of an arbitrary (that is, not-necessarily compartmental) nth order asymptotically stable complex-valued system subjected to additive white noise disturbances. Specifically, we rearrange the elements of the $n \times n$ second-moment matrix into an n^2-dimensional vector whose first n components are the diagonal elements of the second-moment matrix, and whose last $n^2 - n$ components are the off-diagonal elements. Our ability to do this is based upon the following crucial fact: *the diagonal elements of a* (*complex*, *Hermitian*) *nonnegative-definite matrix are* (*real and*) *nonnegative*.[2]

The central result of § 3 is the derivation of an explicit equation that governs the evolution of the diagonal elements of the second-moment equation. As can be seen from (3.18), this equation involves the unusual nonnegative matrix coefficient $e^{Ft} \circ e^{\bar{F}t}$, where F is the dynamics matrix of the original arbitrary (not-necessarily compartmental) state space system, and "$\circ$" denotes Hadamard (Schur) product. Since this system has dynamics that are more complex than an nth order state space system, we confine our attention in § 4 to the steady-state energy distribution.

The goal of § 4 is to determine an nth order state space system whose steady-state solution coincides with the steady-state limit of the nonnegative diagonal system. This requirement leads to a derived dynamics matrix (see (4.12)) involving the elements of the dynamics matrix of the original state space model. To prove that the induced model is asymptotically stable, we consider the case of weak off-diagonal coupling which leads to an M-matrix condition and implies asymptotic stability. A final scaling of the nonnegative system in § 5 shows that in this case the derived model is, in fact, a compartmental model. Finally, notation and identities involving Kronecker and Hadamard products appear in the Appendix.

Notation.

$\mathbb{E}$	expectation
$\mathbb{R}$, $\mathbb{C}$	real field, complex field
$\mathbb{R}^{r \times s}$, $\mathbb{C}^{r \times s}$	$r \times s$ real, complex matrices
I_r or I	$r \times r$ identity matrix
j	$\sqrt{-1}$
A_{kl}	(k, l)-element of $A \in \mathbb{C}^{r \times s}$ (or a subblock of A)
Re A, Im A	real, imaginary part of $A \in \mathbb{C}^{r \times s}$
$\bar{A}$, A^T, A^*	conjugate, transpose, complex conjugate transpose
$\mathcal{R}(A)$, $\mathcal{N}(A)$	range and null space of $A \in \mathbb{R}^{r \times s}$
$A \geqq\geqq 0$	$A \in \mathbb{R}^{r \times s}$ is a nonnegative matrix
$A \circ B$	Hadamard (Schur) (element-by-element) product
$\mathbf{e}$	$[1 \quad 1 \cdots 1]^T$ (boldface distinguishes from exponential)
col$_i(A)$	ith column of A
e_i	ith column of I_r
$\mathcal{E}_{ij}$, $I_{/i}$	see Appendix
diag $(a_1, \ldots, a_n)$	$n \times n$ matrix with diagonal elements $a_1, \ldots, a_n$
$\{A\}$, $\langle A \rangle$	diagonal, off-diagonal part of $A \in \mathbb{C}^{r \times r}$ (see Appendix)
$\otimes$, $\oplus$, vec, vecd, veco	see Appendix

[1] Throughout the paper a nonnegative-definite matrix is assumed to be Hermitian.

[2] Since this paper uses both nonnegative matrices and nonnegative-definite matrices in close proximity, care must be taken to note the distinction.

2. Analysis of compartmental models. To begin we consider a system comprised of compartments or subsystems that interact by exchanging some quantity such as mass, energy, fluid, etc. We shall use energy and power analogies for generic terminology. By applying conservation of energy, energy flow among subsystems and the external environment as shown in Fig. 1 leads to the energy balance equation

$$\dot{E}_i(t) = -\sigma_{ii}E_i(t) + \sum_{\substack{j=1\\ j\neq i}}^{n} \Pi_{ij}(t) + P_i(t), \quad t \geqq 0, \quad i = 1, \ldots, n, \tag{2.1}$$

where, for $i = 1, \ldots, n$,

$E_i(t)$ = energy of the ith subsystem,

σ_{ii} = loss coefficient of the ith subsystem, $\sigma_{ii} \geqq 0$,

$P_i(t)$ = external power applied to the ith subsystem, $P_i(t) \geqq\geqq 0$, $t \geqq 0$,

$\Pi_{ij}(t)$ = net energy flow from the jth subsystem to the ith subsystem, $j \neq i$.

As depicted in Fig. 1, it is assumed that $\Pi_{ij}(t)$ is of the form

$$\Pi_{ij}(t) = \sigma_{ij}E_j(t) - \sigma_{ji}E_i(t), \qquad t \geqq 0, \tag{2.2}$$

where $\sigma_{ij} \geqq 0$, $i \neq j$, $i, j = 1, \ldots, n$. Note that $\Pi_{ji}(t) = -\Pi_{ij}(t)$, $t \geqq 0$. Assembling (2.1) and (2.2) into matrix form yields the overall systems model

$$\dot{E}(t) = AE(t) + P(t), \qquad t \geqq 0, \tag{2.3}$$

where

$$E(t) \triangleq [E_1(t) \cdots E_n(t)]^T, \qquad P(t) \triangleq [P_1(t) \cdots P_n(t)]^T,$$

and $A = [A_{ij}]_{i,j=1}^{n}$ is defined by

$$A_{ii} \triangleq -\sum_{j=1}^{n} \sigma_{ji}, \qquad i = 1, \ldots, n, \tag{2.4}$$

$$A_{ij} \triangleq \sigma_{ij}, \quad i \neq j, \quad i, j = 1, \ldots, n. \tag{2.5}$$

Letting $\sigma \triangleq [\sigma_{ij}]_{i,j=1}^{n}$ and using the matrix operators introduced in the Appendix, the matrix A can be written compactly as

$$A = -\mathrm{vecd}^{-1}(\sigma^T \mathbf{e}) + \langle \sigma \rangle.$$

As shown in the Appendix, the operator "vecd" extracts the diagonal elements of a matrix to form a column vector, while "vecd^{-1}" transforms a column vector into a

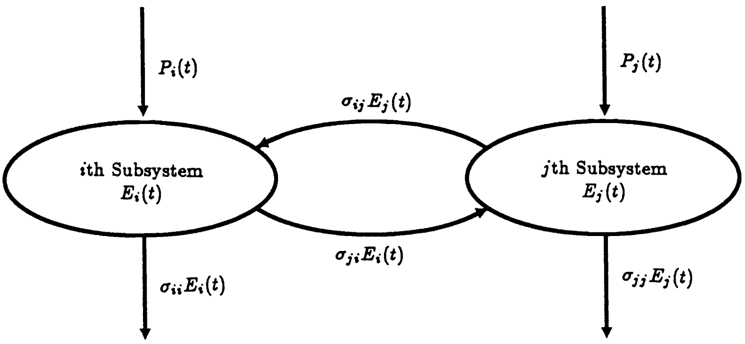


FIG. 1. *Compartmental model involving interconnected subsystems.*

diagonal matrix. Furthermore, $\langle\sigma\rangle$ denotes the off-diagonal matrix comprised of only the off-diagonal elements of σ with the diagonal elements replaced by zeros.

An important special form of (2.2) arises when $\sigma_{ij} = \sigma_{ji}$ for some $i \neq j$. In this case $\Pi_{ij}(t)$ can be written as

$$\Pi_{ij}(t) = \sigma_{ij}[E_j(t) - E_i(t)], \quad t \geqq 0, \tag{2.6}$$

which can be interpreted thermodynamically as saying that heat flow is proportional to temperature difference. Note that A is symmetric if and only if σ is symmetric.

The solution $E(t)$ to (2.3) can be written explicitly as

$$E(t) = e^{At}E(0) + \int_0^t e^{A(t-s)}P(s)\,ds, \tag{2.7}$$

where the function $P(\cdot)$ is assumed to be such that the integral in (2.7) exists. To analyze (2.7), we begin by noting that A is *essentially nonnegative* [13], that is, the off-diagonal elements of A are nonnegative. Equivalently, $-A$ is a *Z-matrix* [6], that is, $-A$ has nonpositive off-diagonal elements. The following lemma concerns the exponential of an essentially nonnegative matrix. Variations of this result appear in [6, p. 146], [13, p. 74], [55, p. 37], and [60, p. 207].

LEMMA 2.1. *Let $B \in \mathbb{R}^{n\times n}$. Then B is essentially nonnegative if and only if e^{Bt} is nonnegative for all $t \geqq 0$.*

Proof. If B is essentially nonnegative, then there exists $\beta > 0$ sufficiently large such that $\hat{B} \triangleq \beta I + B$ is nonnegative. Consequently, $e^{\hat{B}t}$ is nonnegative for all $t \geqq 0$, and thus $e^{Bt} = e^{-\beta t}e^{\hat{B}t}$ is nonnegative for all $t \geqq 0$. Conversely, suppose that $B_{ij} < 0$ for some $i \neq j$. Then, since $(e^{Bt})_{ij} = tB_{ij} + O(t^2)$ as $t \to 0$ for $i \neq j$, it follows that $(e^{Bt})_{ij} < 0$ for some $t > 0$ sufficiently small. Hence e^{Bt} is not nonnegative for all $t \geqq 0$. □

Since A is essentially nonnegative, its exponential is nonnegative on $[0, \infty)$. If $E(0)$ and $P(t)$ are both nonnegative, then it follows immediately from (2.7) that $E(t)$ is nonnegative.

PROPOSITION 2.1. *Suppose that $E(0) \geqq\geqq 0$ and $P(t) \geqq\geqq 0$, $t \geqq 0$. Then the solution $E(t)$ to (2.3) is nonnegative for all $t \geqq 0$.*

Henceforth we focus on the case in which the externally applied power $P(t)$ is constant, that is, $P(t) \equiv P$. In this case (2.3) and (2.7) become

$$\dot{E}(t) = AE(t) + P, \qquad t \geqq 0, \tag{2.8}$$

and

$$E(t) = e^{At}E(0) + \int_0^t e^{As}\,dsP, \qquad t \geqq 0. \tag{2.9}$$

The following lemma summarizes several properties of A that are useful in analyzing (2.9). Recall [61] that the *index k* of a real matrix M, denoted ind (M), is defined to be the smallest nonnegative integer k such that rank M^k = rank M^{k+1}. (Here $M^0 \triangleq I$.) Equivalently, ind (M) is the size of the largest Jordan block of M associated with the eigenvalue zero. Furthermore, recall that if ind $(M) \leqq 1$, then the Drazin inverse M^D specializes to the group inverse $M^{\#}$ of M. It can be seen that ind $(M) \leqq 1$ and every eigenvalue of M either has negative real part or is zero if and only if $\lim_{t\to\infty} e^{Mt}$ exists. In this case M is called *semistable*. Finally, let $\mathcal{R}(M)$ and $\mathcal{N}(M)$ denote the range and nullspace of M, respectively.

LEMMA 2.2. *The matrix A defined by* (2.4), (2.5) *has the following properties*:

(i) $-A$ is an M-matrix,

(ii) If λ is an eigenvalue of A then either $\operatorname{Re}\lambda < 0$ or $\lambda = 0$,

(iii) $\operatorname{ind}(A) \leqq 1$,

(iv) A is semistable, and $\lim_{t\to\infty} e^{At} = I - AA^{\#} \geqq\geqq 0$,

(v) $\mathscr{R}(A) = \mathcal{N}(I - AA^{\#}),\ \mathcal{N}(A) = \mathscr{R}(I - AA^{\#})$,

(vi) $\int_0^t e^{As}\,ds = A^{\#}(e^{At} - I) + (I - AA^{\#})t,\ t \geqq 0$,

(vii) $\int_0^\infty e^{As}\,dsP$ exists if and only if $P \in \mathscr{R}(A)$,

(viii) If $P \in \mathscr{R}(A)$, then $\int_0^\infty e^{As}\,dsP = -A^{\#}P$,

(ix) If $P \in \mathscr{R}(A)$ and $P \geqq\geqq 0$, then $-A^{\#}P \geqq\geqq 0$,

(x) A is nonsingular if and only if $-A$ is a nonsingular M-matrix,

(xi) If A is nonsingular, then A is asymptotically stable and $-A^{-1} \geqq\geqq 0$.

Proof. Since $-A^T\mathbf{e} \geqq\geqq 0$ and $-A$ is a Z-matrix, it follows from [62, Thm. 1, p. 237] or [6, Exercise 6.4.14, p. 155] that $-A^T$, and hence $-A$, is an M-matrix with "property c" (see [6, Def. 6.4.10, p. 152]), which proves (i). Since $-A$ is an M-matrix, it follows from [6, Prop. (E_{11}), p. 150] that the real part of each nonzero eigenvalue of A is negative, which proves (ii). From [6, Lemma 6.4.11, p. 153], it follows that $\operatorname{ind}(A) \leqq 1$, thus proving (iii). To prove (iv), write $A = S[\begin{smallmatrix}A_0 & 0\\ 0 & 0\end{smallmatrix}]S^{-1}$, where A_0 is asymptotically stable. Then

$$e^{At} = S\begin{bmatrix} e^{A_0t} & 0 \\ 0 & I \end{bmatrix}S^{-1} \to S\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}S^{-1} = I - AA^{\#} \quad \text{as } t \to \infty,$$

which proves (iv). Note that $I - AA^{\#}$ is nonnegative since e^{At} is nonnegative for all $t \geqq 0$. To prove (v), note that if $(I - AA^{\#})x = 0$, then $x = AA^{\#}x \in \mathscr{R}(A)$. Conversely, if $x \in \mathscr{R}(A)$, then there exists $y \in \mathbb{R}^n$ such that $x = Ay$ so that $AA^{\#}x = AA^{\#}Ay = Ay = x$. The second identity follows similarly. Next, (vi) follows from [61, Thm. 9.2.4] and can be verified directly. Statement (vii) is a direct consequence of (v) and (vi), while (viii) follows from (iv) and (vi). Next, (ix) follows from (viii) and the fact that $e^{At} \geqq\geqq 0$, $t \geqq 0$. Finally, (x) follows from (i) or [6, p. 137], and (xi) follows from (ii) and (ix) with $P = e_i$, $i = 1, \ldots, n$. □

Remark 2.1. Properties (ii) and (iii) imply that the homogeneous system $\dot{E}(t) = AE(t)$ is stable in the sense of Lyapunov. This result is given by [62, Thm. 1] in terms of the set $\mathscr{W}\mathscr{L}$. The same result is given by [3, Lemmas 1 and 2] and is attributed to [1]. The result (ix) that $x \in \mathscr{R}(A)$ and $x \geqq\geqq 0$ imply that $-A^{\#}x \geqq\geqq 0$ is given by [63, Thm. 3].

By using Lemma 2.2 we can obtain an expression for the steady-state energy distribution $\lim_{t\to\infty} E(t)$. For notational convenience we denote this limit simply by E.

PROPOSITION 2.2. *Suppose that* $E(0) \geqq\geqq 0$ *and* $P \geqq\geqq 0$ *and let* $E(t)$ *be given by* (2.9). *Then* $E \triangleq \lim_{t\to\infty} E(t)$ *exists if and only if* $P \in \mathscr{R}(A)$. *In this case E is given by*

$$E = (I - AA^{\#})E(0) - A^{\#}P, \tag{2.10}$$

and $E \geqq\geqq 0$. If, in addition, A is nonsingular, then E exists for all $P \geqq\geqq 0$ and is given by

$$E = -A^{-1}P. \tag{2.11}$$

In equilibrium, the dynamic system (2.8) becomes

$$0 = AE + P. \tag{2.12}$$

We now show that the steady-state solution (2.10) is in fact an equilibrium solution to (2.8) and, furthermore, all solutions to (2.12) are of the form (2.10).

PROPOSITION 2.3. *Let $P \in \mathbb{R}^n$. Then* (2.12) *has a solution $E \in \mathbb{R}^n$ if and only if $P = AA^\# P$. Furthermore, $E \in \mathbb{R}^n$ is a solution to* (2.12) *if and only if there exists $E(0) \in \mathbb{R}^n$ such that* (2.10) *is satisfied.*

Proof. Clearly, (2.12) has a solution $E \in \mathbb{R}^n$ if and only if $P \in \mathcal{R}(A)$. By (v) of Lemma 2.2, $P \in \mathcal{R}(A)$ if and only if $P \in \mathcal{N}(I - AA^\#)$, that is, $P = AA^\# P$. Next, it is easy to verify that E given by (2.10) is a solution to (2.12). Conversely, if E satisfies (2.12) then $z \triangleq E + A^\# P$ is in the null space of A. Since by (v) of Lemma 2.2 the null space of A coincides with the range of $I - AA^\#$, it follows that there exists $E(0) \in \mathbb{R}^n$ such that $z = (I - AA^\#)E(0)$, which yields (2.10). □

Remark 2.2. Proposition 2.3 is entirely analogous to the standard result involving the Moore–Penrose generalized inverse (see, for example, [64, p. 37]). Except for the necessity of the second statement, the result is given by [65, Lemma 5.1].

Writing $E = [E_1 \cdots E_n]^T$ and $P = [P_1 \cdots P_n]^T$, the ith component of (2.12) can be written as

$$0 = -\sigma_{ii}E_i + \sum_{\substack{j=1\\ j\neq i}}^{n} [\sigma_{ij}E_j - \sigma_{ji}E_i] + P_i, \tag{2.13}$$

which can be viewed as an energy balance relation.

When the rank of A is equal to $n - 1$, it is possible to simplify expression (2.10). The following lemma will be useful.

LEMMA 2.3. *Suppose rank $A = n - 1$ and let $v \in \mathbb{R}^n$ satisfy $Av = 0$. Then either $v \geqq\geqq 0$* or *$-v \geqq\geqq 0$.*

Proof. Since $v \in \mathcal{N}(A)$, it follows from (v) of Lemma 2.2 that $v \in \mathcal{R}(I - AA^\#)$. Since rank $A = n - 1$, it follows that $\mathcal{N}(A)$ is one dimensional and thus rank $(I - AA^\#) = 1$. By (iv) of Lemma 2.2, $I - AA^\#$ is also nonnegative. Since $I - AA^\#$ is also nonzero, there exists a nonnegative vector w such that $(I - AA^\#)w$ is also nonzero (and nonnegative). Since v and $(I - AA^\#)w$ both lie in the same one-dimensional subspace, there exists $\beta \in \mathbb{R}$ such that $v = \beta(I - AA^\#)w$. If $\beta \geqq 0$ then $v \geqq\geqq 0$, whereas if $\beta \leqq 0$ then $v \leqq\leqq 0$. □

COROLLARY 2.1. *Suppose that $E(0) \geqq\geqq 0$, $P \geqq\geqq 0$, and $P \in \mathcal{R}(A)$. Furthermore, assume that rank $A = n - 1$ and let $v \in \mathbb{R}^n$, $v \neq 0$, $v \geqq\geqq 0$ satisfy $Av = 0$. Then the steady-state energy distribution E given by* (2.10) *has the form*

$$E = \beta v - A^\# P, \tag{2.14}$$

where $\beta = \|(I - AA^\#)E(0)\| / \|v\|$ and $\|\cdot\|$ denotes an arbitrary norm on $\mathbb{R}^n$.

Proof. Since A is singular there exists nonzero $v \in \mathcal{N}(A)$. Furthermore, since rank $A = n - 1$, it follows from Lemma 2.3 that either $v \geqq\geqq 0$ or $-v \geqq\geqq 0$. Without loss of generality, let v be chosen such that $v \geqq\geqq 0$. Since rank $A = n - 1$, it follows that $\mathcal{N}(A)$ and thus $\mathcal{R}(I - AA^\#)$ are one dimensional. Thus there exists $\beta \geqq 0$ such that $\beta v = (I - AA^\#)E(0)$. Note that β is necessarily nonnegative since v is nonzero and v and $(I - AA^\#)E(0)$ are both nonnegative vectors. Taking norms yields the given expression for β. □

Remark 2.3. A sufficient condition for a singular M-matrix A to have rank $n - 1$ is for A to be irreducible (see [6, Thm. 6.4.16, p. 156]). In this case the nonnegative vector $v \in \mathcal{N}(A)$ actually has all positive components (see [6]).

As an application of Corollary 2.1 we consider the case in which $\sigma_{ii} = 0$, $i = 1, \ldots, n$. Then $\mathbf{e}^T A = 0$, which implies that rank $A \leqq n - 1$. In this case it is easy to see that when $P = 0$ the total system energy is conserved, since (2.8) implies $\mathbf{e}^T\dot{E}(t) = \mathbf{e}^T AE(t) = 0$. If we also assume that $A\mathbf{e} = 0$, we obtain the following result.

COROLLARY 2.2. *Suppose that* $\sigma_{ii} = 0$, $i = 1, \ldots, n$, $A\mathbf{e} = 0$, *and* rank $A = n - 1$. *If* $E(0) \geqq\geqq 0$ *and* $P = 0$, *then the steady-state energy distribution* E *given by* (2.10) *has the form*

$$E = \left[\frac{1}{n} \sum_{i=1}^{n} E_i(0)\right]\mathbf{e}. \tag{2.15}$$

Proof. Since $A\mathbf{e} = 0$, it follows from Corollary 2.1 and (2.14) that $E = \beta\mathbf{e}$, where (choosing the Euclidean norm in Corollary 2.1) $\beta = n^{-1/2}[E^T(0)(I - AA^{\#})^T(I - AA^{\#})E(0)]^{1/2}$. Next, since $A\mathbf{e} = A^T\mathbf{e} = 0$ and rank $A = n - 1$, it follows that A is an EP matrix ([61, p. 74]). Consequently, $AA^{\#}$ is symmetric and, in particular, $I - AA^{\#} = (1/n)\mathbf{e}\mathbf{e}^T$. This implies that $\beta = n^{-1}\mathbf{e}^T E(0)$, which yields (2.15). □

The result (2.15) shows that under the stated assumptions each component of the steady-state energy vector is equal, that is, the steady-state energy is uniformly distributed over all states. This phenomenon is known as *equipartition of energy* [34], [41], [50].

Henceforth we consider the case in which A is nonsingular, that is, in which $-A$ is a nonsingular M-matrix. Numerous necessary and sufficient conditions for a Z-matrix to be a nonsingular M-matrix are given by [6, Thm. 6.2.3]. The following easily verified condition is sufficient but not necessary.

LEMMA 2.4. *If* $\sigma_{ii} > 0$, $i = 1, \ldots, n$, *then* $-A$ *is a nonsingular M-matrix.*

Proof. If $\sigma_{ii} > 0$, $i = 1, \ldots, n$, then $-A^T\mathbf{e} = [\sigma_{11} \cdots \sigma_{nn}]^T \gg 0$. By [6, Condition ($I_{27}$), p. 136], $-A^T$, and hence $-A$, is a nonsingular M-matrix. □

Next we examine the steady-state energy distribution E as a function of the coupling parameters σ_{ij}, $i \neq j$. Specifically, we wish to determine the steady-state energy distribution E in the limit of strong coupling, that is, as $\sigma_{ij} \to \infty$, $i \neq j$. To do this, define

$$A_\alpha \triangleq -D + \alpha C,$$

where the diagonal matrix D is defined by $D = \operatorname{diag}(\sigma_{11}, \ldots, \sigma_{nn})$ and the matrix C of coupling parameters is defined by

$$C_{ii} = -\sum_{\substack{j=1\\ j\neq i}}^{n} \sigma_{ji}, \qquad i = 1, \ldots, n, \tag{2.16}$$

$$C_{ij} = \sigma_{ij}, \quad i \neq j, \quad i, j = 1, \ldots, n. \tag{2.17}$$

In the notation of the Appendix,

$$D = \{\sigma\}, \qquad C = -\operatorname{vecd}^{-1}(\langle\sigma\rangle^T\mathbf{e}) + \langle\sigma\rangle = A + D, \tag{2.18}$$

where $\{\sigma\}$ and $\langle\sigma\rangle$ denote the diagonal and off-diagonal portions, respectively, of A. Note that $A = A_1$. Furthermore, note that if $\sigma_{ii} > 0$, $i = 1, \ldots, n$, then $-A_\alpha$ is a nonsingular M-matrix for all $\alpha \geqq 0$. For $P \geqq\geqq 0$, let E_α denote the steady-state energy distribution with A replaced by A_α, that is,

$$E_\alpha = -A_\alpha^{-1}P. \tag{2.19}$$

Note that letting $\alpha \to \infty$ corresponds to letting $\sigma_{ij} \to \infty$, $i \neq j$. The following result provides an expression for $\lim_{\alpha\to\infty} E_\alpha$, which is the steady-state energy distribution in the limit of strong coupling.

PROPOSITION 2.4. *If $\sigma_{ii} > 0$, $i = 1, \dots, n$, then $E_\infty \triangleq \lim_{\alpha\to\infty} E_\alpha$ exists and is given by*

$$E_\infty = [D^{-1} - D^{-1}CD^{-1}(CD^{-1})^{\#}]P. \tag{2.20}$$

Proof. Since $\mathbf{e}^T CD^{-1} = 0$, it follows from [6, Exercise 6.4.14, p. 155] that $-CD^{-1}$ is a singular M-matrix with "property c." Hence [6, Lemma 6.4.11, p. 153] implies that ind $(CD^{-1}) = 1$. From [61, Cor. 7.6.4], we obtain

$$\begin{aligned} E_\infty &= \lim_{\alpha\to\infty} [D - \alpha C]^{-1}P \\ &= D^{-1} \lim_{\beta\to 0} \beta[\beta I - CD^{-1}]^{-1}P \\ &= [D^{-1} - D^{-1}CD^{-1}(CD^{-1})^{\#}]P. \qquad \square \end{aligned}$$

A minor variation of the proof of Proposition 2.4 shows that E_∞ can be written equivalently as either

$$E_\infty = [D^{-1} - D^{-1}CD^{-1/2}(D^{-1/2}CD^{-1/2})^{\#}D^{-1/2}]P \tag{2.21}$$

or

$$E_\infty = [D^{-1} - D^{-1}C(D^{-1}C)^{\#}D^{-1}]P. \tag{2.22}$$

The symmetry of the expression (2.21) will be useful in obtaining a more explicit expression for E_∞ when C is symmetric and has rank $n-1$. The next result shows that in this case strong coupling leads to energy equipartition.

COROLLARY 2.3. *Assume that $\sigma_{ii} > 0$, $i = 1, \dots, n$, and suppose that C is symmetric and rank $C = n - 1$. Then E_∞ is given by*

$$E_\infty = \left(\frac{\mathbf{e}^T P}{\mathbf{e}^T D\mathbf{e}}\right)\mathbf{e}. \tag{2.23}$$

Proof. For convenience define $\hat{C} \triangleq D^{-1/2}CD^{-1/2}$, which also has rank $n-1$. Since C and $\hat{C}$ are symmetric, it follows that $\hat{C}D^{1/2}\mathbf{e} = 0$. By decomposing $\hat{C}$ it can be shown that $I - \hat{C}\hat{C}^{\#} = (\mathbf{e}^T D\mathbf{e})^{-1}D^{1/2}\mathbf{e}\mathbf{e}^T D^{1/2}$. Hence, using (2.21), we obtain

$$\begin{aligned} E_\infty &= D^{-1/2}[I - \hat{C}\hat{C}^{\#}]D^{-1/2}P \\ &= (\mathbf{e}^T D\mathbf{e})^{-1}D^{-1/2}D^{1/2}\mathbf{e}\mathbf{e}^T D^{1/2}D^{-1/2}P \\ &= \left(\frac{\mathbf{e}^T P}{\mathbf{e}^T D\mathbf{e}}\right)\mathbf{e}. \qquad \square \end{aligned}$$

3. Second-moment analysis of state space systems.

In this section we consider an arbitrary asymptotically stable linear system subjected to additive white noise disturbances. The second-moment matrix of the state then satisfies a matrix Lyapunov differential equation. From this matrix differential equation, we then extract a vector differential equation for the diagonal elements of the second-moment matrix. These diagonal elements are the second moments of the individual states. This section relies heavily on Kronecker matrix algebra, which is summarized in the Appendix.

To begin, consider the state space differential equation

$$\dot{x}(t) = Fx(t) + Gw(t), \qquad t \geqq 0, \tag{3.1}$$

where $F \in \mathbb{C}^{n\times n}$, $G \in \mathbb{C}^{n\times d}$, $w(\cdot)$ is d-dimensional zero-mean stationary Gaussian white noise with intensity I_d, and $x(0)$ is Gaussian distributed with not-necessarily zero mean.

The second-moment matrix of $x(t)$ defined by $Q(t) \triangleq \mathbb{E}[x(t)x^*(t)] \in \mathbb{C}^{n\times n}$ satisfies the matrix Lyapunov differential equation [66, p. 101],

$$(3.2)\qquad \dot{Q}(t) = FQ(t) + Q(t)F^* + V, \qquad t \geqq 0,$$

where $Q(0) = \mathbb{E}[x(0)x^*(0)]$ and $V \triangleq GG^*$. The solution $Q(t)$ to (3.2) is given explicitly by

$$(3.3)\qquad Q(t) = e^{Ft}Q(0)e^{F^*t} + \int_0^t e^{Fs}Ve^{F^*s}\,ds, \qquad t \geqq 0,$$

which shows that $Q(t)$ is a (Hermitian) nonnegative-definite matrix.

Applying the vec operator [67], [68] to (3.2) and using (A.7) (see the Appendix) yields

$$(3.4)\qquad \operatorname{vec} \dot{Q}(t) = (\bar{F} \oplus F) \operatorname{vec} Q(t) + \operatorname{vec} V.$$

Next, we use the $n^2 \times n^2$ orthogonal permutation matrix U to rearrange the components of vec $Q(t)$ so that the diagonal elements of $Q(t)$ appear as the first n elements. Hence (3.4) implies

$$(3.5)\qquad U \operatorname{vec} \dot{Q}(t) = MU \operatorname{vec} Q(t) + U \operatorname{vec} V,$$

where M is defined by

$$(3.6)\qquad M \triangleq U(\bar{F} \oplus F)U^{-1} = U(\bar{F} \oplus F)U^T.$$

Since $U = [\begin{smallmatrix} U_d \\ U_o \end{smallmatrix}]$, identities (A.13) and (A.14) imply

$$U \operatorname{vec} Q(t) = \begin{bmatrix} U_d \operatorname{vec} Q(t) \\ U_o \operatorname{vec} Q(t) \end{bmatrix} = \begin{bmatrix} \operatorname{vecd} Q(t) \\ \operatorname{veco} Q(t) \end{bmatrix},$$

and similarly for vec $\dot{Q}(t)$ and vec V.

Next, defining

$$E(t) \triangleq \operatorname{vecd} Q(t), \qquad P \triangleq \operatorname{vecd} V,$$
$$\hat{E}(t) \triangleq \operatorname{veco} Q(t), \qquad \hat{P} \triangleq \operatorname{veco} V,$$

(3.5) can be written as

$$(3.7)\qquad \begin{bmatrix} \dot{E}(t) \\ \dot{\hat{E}}(t) \end{bmatrix} = M \begin{bmatrix} E(t) \\ \hat{E}(t) \end{bmatrix} + \begin{bmatrix} P \\ \hat{P} \end{bmatrix}.$$

Note that $E(t)$ and P are real and nonnegative, whereas $\hat{E}(t)$ and $\hat{P}$ are generally complex. Furthermore, $Q(t)$ and V can be reconstructed from $E(t)$, $\hat{E}(t)$, P, and $\hat{P}$ by means of

$$(3.8)\quad Q(t) = \operatorname{vecd}^{-1}(E(t)) + \operatorname{veco}^{-1}(\hat{E}(t)), \qquad V = \operatorname{vecd}^{-1}(P) + \operatorname{veco}^{-1}(\hat{P}).$$

Next, partition M defined by (3.6) as

$$(3.9)\qquad M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix},$$

where, using (3.6), the subblocks of M are given by

$$M_{11} \triangleq U_d(\bar{F} \oplus F)U_d^T, \qquad M_{12} \triangleq U_d(\bar{F} \oplus F)U_o^T,$$
$$M_{21} \triangleq U_o(\bar{F} \oplus F)U_d^T, \qquad M_{22} \triangleq U_o(\bar{F} \oplus F)U_o^T.$$

Using (A.11) and (A.17)–(A.20), M_{11}, M_{12}, M_{21}, and M_{22} can be simplified somewhat as

$$M_{11} = 2\operatorname{Re}\{F\}, \qquad M_{12} = U_d(\langle \bar{F}\rangle \oplus \langle F\rangle)U_o^T, \tag{3.10}$$

$$M_{21} = U_o(\langle \bar{F}\rangle \oplus \langle F\rangle)U_d^T, \qquad M_{22} = U_o(\{\bar{F}\} \oplus \{F\})U_o^T + U_o(\langle \bar{F}\rangle \oplus \langle F\rangle)U_o^T, \tag{3.11}$$

where $\{F\}$ and $\langle F\rangle$ denote the diagonal and off-diagonal portions of F, respectively. Note that M_{11} is real and diagonal and that the term $U_o(\{\bar{F}\} \oplus \{F\})U_o^T$ appearing in M_{22} is diagonal.

We now wish to eliminate $\hat{E}(t)$ from (3.7) to obtain an equation solely in terms of $E(t)$. Hence solving for $\hat{E}(t)$ and substituting this expression into the equation for $\dot{E}(t)$ with $\hat{E}(0) = 0$ and $\hat{P} = 0$ yields the integro-differential equation

$$\dot{E}(t) = M_{11}E(t) + M_{12}\int_0^t e^{M_{22}(t-s)}M_{21}E(s)\,ds + P. \tag{3.12}$$

Again, using (3.7), it follows that the solution to (3.12) is given explicitly by

$$E(t) = [I_n \quad 0]e^{Mt}\begin{bmatrix} I_n \\ 0 \end{bmatrix}E(0) + [I_n \quad 0]\int_0^t e^{Ms}\,ds\begin{bmatrix} I_n \\ 0 \end{bmatrix}P. \tag{3.13}$$

Since $E(t)$ is nonnegative if $E(0)$ and P are nonnegative, we expect the coefficients of $E(0)$ and P in (3.13) to be nonnegative matrices. This is illustrated by the following result which provides explicit expressions for these matrices.

PROPOSITION 3.1. *The following identities are satisfied*:

$$[I_n \quad 0]e^{Mt}\begin{bmatrix} I_n \\ 0 \end{bmatrix} = e^{Ft} \circ e^{\bar{F}t}, \tag{3.14}$$

$$[I_n \quad 0]\int_0^t e^{Ms}\,ds\begin{bmatrix} I_n \\ 0 \end{bmatrix} = \int_0^t e^{Fs} \circ e^{\bar{F}s}\,ds. \tag{3.15}$$

Proof. From (3.6) and (A.8) it follows that

$$\begin{aligned}[I_n \quad 0]e^{Mt}\begin{bmatrix} I_n \\ 0 \end{bmatrix} &= [I_n \quad 0]Ue^{(\bar{F}\oplus F)t}U^T\begin{bmatrix} I_n \\ 0 \end{bmatrix} \\ &= U_d e^{(\bar{F}\oplus F)t}U_d^T \\ &= U_d(e^{\bar{F}t} \otimes e^{Ft})U_d^T \\ &= e^{\bar{F}t} \circ e^{Ft},\end{aligned}$$

where the last step follows from (A.11). Integrating (3.14) over $[0, t)$ yields (3.15). □

From (3.14), we obtain the following result.

COROLLARY 3.1. *The following identities are satisfied*:

$$M_{12}M_{21} = 2F \circ \bar{F} + 2\operatorname{Re}\{F^2\} - 4(\operatorname{Re}\{F\})^2, \tag{3.16}$$

$$\begin{aligned} M_{12}M_{22}M_{21} &= 2\operatorname{Re}\{F^3\} - 8\operatorname{Re}\{F\}\operatorname{Re}\{F^2\} + 8(\operatorname{Re}\{F\})^3 \\ &\quad + 6\operatorname{Re}(F \circ \bar{F}^2) - 4\operatorname{Re}\{F\}(F \circ \bar{F}) - 4(F \circ \bar{F})\operatorname{Re}\{F\}. \end{aligned} \tag{3.17}$$

Proof. First note that for $t \to 0$

$$[I_n \quad 0]e^{Mt}\begin{bmatrix} I_n \\ 0 \end{bmatrix} = I + M_{11}t + \tfrac{1}{2}(M_{11}^2 + M_{12}M_{21})t^2$$
$$+ \tfrac{1}{6}(M_{11}^3 + M_{11}M_{12}M_{21} + M_{12}M_{21}M_{11} + M_{12}M_{22}M_{21})t^3 + O(t^4),$$

while

$$e^{Ft} \circ e^{\bar{F}t} = (I + Ft + \tfrac{1}{2}F^2t^2 + \tfrac{1}{6}F^3t^3 + O(t^4)) \circ (I + \bar{F}t + \tfrac{1}{2}\bar{F}^2t^2 + \tfrac{1}{6}\bar{F}^3t^3 + O(t^4))$$
$$= I + \{F + \bar{F}\}t + (\tfrac{1}{2}\{F^2 + \bar{F}^2\} + F \circ \bar{F})t^2$$
$$+ (\tfrac{1}{3}\operatorname{Re}\{F^3\} + \operatorname{Re}(F \circ \bar{F}^2))t^3 + O(t^4).$$

Equating terms of $O(t^2)$ and $O(t^3)$ yields (3.16) and (3.17). □

COROLLARY 3.2. *The matrix $M_{12}M_{21}$ is essentially nonnegative.*

Proof. The result follows from the fact that $\langle M_{12}M_{21}\rangle = 2\langle F \circ \bar{F}\rangle \geqq\geqq 0$. □

Using expressions (3.14) and (3.15), (3.13) for $E(t)$ can be written as

$$E(t) = e^{Ft} \circ e^{\bar{F}t}E(0) + \int_0^t e^{Fs} \circ e^{\bar{F}s}\,ds P. \tag{3.18}$$

For comparison, let us recall the solution (2.9) to the compartmental model (2.8) given by

$$E(t) = e^{At}E(0) + \int_0^t e^{As}\,ds P. \tag{3.19}$$

These models, that is, (3.18) and (3.19), will coincide if and only if

$$e^{At} = e^{Ft} \circ e^{\bar{F}t}, \qquad t \geqq 0. \tag{3.20}$$

In general, however, there does not exist a matrix A satisfying (3.19) for the obvious reason that $e^{Ft} \circ e^{\bar{F}t}$ involves more spectral content than can be provided by the exponential of a single $n \times n$ matrix. It is easy to see that there exists a matrix A such that (3.19) and (3.20) coincide if and only if F is diagonal, in which case $A = F + \bar{F}$. To proceed, let us consider the steady-state problem. To guarantee that $\lim_{t\to\infty} E(t)$ exists, we shall assume that F is semistable. The following result will be useful.

LEMMA 3.1. *If F is semistable, then M is semistable.*

Proof. Since $e^{Mt} = U(e^{\bar{F}t} \otimes e^{Ft})U^T$, the existence of $\lim_{t\to\infty} e^{Ft}$ implies the existence of $\lim_{t\to\infty} e^{Mt}$. □

PROPOSITION 3.2. *Suppose that F is semistable and assume that $P \in \mathcal{N}(I_n - U_d(\bar{F} \oplus F)(\bar{F} \oplus F)^{\#}U_d^T)$. Then $E \triangleq \lim_{t\to\infty} E(t)$ exists and is given by*

$$E = (I - FF^{\#}) \circ (I - \bar{F}\bar{F}^{\#})E(0) - U_d(\bar{F} \oplus F)^{\#}U_d^T P. \tag{3.21}$$

Proof. The first term in (3.21) follows from (iv) of Lemma 2.2. Since ind $(M) = 1$, we have (using (vi) of Lemma 2.2)

$$\int_0^t e^{Fs} \circ e^{\bar{F}s}\,ds P = [I_n \quad 0]\int_0^t e^{Ms}\,ds\begin{bmatrix} I_n \\ 0 \end{bmatrix}P$$
$$= [I_n \quad 0]M^{\#}(e^{Mt} - I)\begin{bmatrix} I_n \\ 0 \end{bmatrix}P + [I_n \quad 0](I - MM^{\#})\begin{bmatrix} I_n \\ 0 \end{bmatrix}Pt.$$

Since $[I_n \quad 0](I - MM^{\#})[{}^{I_n}_{0}] = I_n - U_d(\bar{F} \oplus F)(\bar{F} \oplus F)^{\#} U_d^T$, the term linear in t is zero by assumption. Consequently, as $t \to \infty$,

$$\int_0^t e^{Fs} \circ e^{\bar{F}s}\, ds P \to -[I_n \quad 0] M^{\#} \begin{bmatrix} I_n \\ 0 \end{bmatrix} P,$$

which is equal to $-U_d(\bar{F} \oplus F)^{\#} U_d^T P$. □

When the rank of F is $n - 1$, some simplification is possible.

COROLLARY 3.3. *Let F and P be as in Proposition 3.2. Furthermore, assume that* rank $F = n - 1$ *and let $v \in \mathbb{C}^n$, $v \neq 0$, satisfy $Fv = 0$. Then E is given by*

$$E = \left(\frac{(v \circ \bar{v})^T E(0)}{(v^* v)^2} \right) v \circ \bar{v} - U_d(\bar{F} \oplus F)^{\#} U_d^T P. \tag{3.22}$$

Proof. Since rank $(I - FF^{\#}) = 1$, it follows that $I - FF^{\#} = wy^*$ for some nonzero $w, y \in \mathbb{C}^n$. Hence $v = (y^* v) w$, $y = (y^* v / v^* v)(I - FF^{\#})^* v$, and $I - FF^{\#} = (v^* v)^{-1} v v^* (I - FF^{\#})$. Using (A.10) in (3.21) now yields (3.22). □

4. Steady-state compartmental modeling of the diagonal system. In the previous section it was shown that the evolution of the diagonal portion of the second-moment matrix cannot generally be modeled by means of an nth-order state space system. Hence we now focus our attention on the steady-state solution to the diagonal system. Our goal is to determine conditions under which the steady-state energy distribution of the diagonal system coincides with the steady-state energy distribution of a compartmental model. Henceforth (and without further notice) we assume that F is asymptotically stable, that is, every eigenvalue of F has negative real part.

Since F is asymptotically stable, $Q \triangleq \lim_{t \to \infty} Q(t)$ exists, is nonnegative definite, and is given by

$$Q = \int_0^{\infty} e^{Fs} V e^{F^* s}\, ds, \tag{4.1}$$

which is the unique solution to the algebraic Lyapunov equation

$$0 = FQ + QF^* + V. \tag{4.2}$$

Note that Q is independent of $Q(0)$. Furthermore, since $\bar{F} \oplus F$ is asymptotically stable [67], [68], M is asymptotically stable so that (3.13) can be written as

$$E(t) = [I_n \quad 0] e^{Mt} \begin{bmatrix} I_n \\ 0 \end{bmatrix} E(0) + [I_n \quad 0] M^{-1}(e^{Mt} - I) \begin{bmatrix} I_n \\ 0 \end{bmatrix} P. \tag{4.3}$$

Now letting $t \to \infty$ in (3.7) yields

$$\begin{bmatrix} E \\ \hat{E} \end{bmatrix} = -M^{-1} \begin{bmatrix} P \\ \hat{P} \end{bmatrix}, \tag{4.4}$$

where $E \triangleq \lim_{t \to \infty} E(t)$ and $\hat{E} \triangleq \lim_{t \to \infty} \hat{E}(t)$. For convenience, partition M^{-1} as

$$M^{-1} = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix}, \tag{4.5}$$

where

$$N_{11} \triangleq U_d(\bar{F} \oplus F)^{-1} U_d^T, \qquad N_{12} \triangleq U_d(\bar{F} \oplus F)^{-1} U_o^T,$$

$$N_{21} \triangleq U_o(\bar{F} \oplus F)^{-1} U_d^T, \qquad N_{22} \triangleq U_o(\bar{F} \oplus F)^{-1} U_o^T.$$

Next, letting $t \rightarrow \infty$ in (4.3) yields

$$N_{11} = -\int_0^\infty e^{Ft} \circ e^{\bar{F}t}\,dt, \tag{4.6}$$

which shows that $-N_{11}$ is a (real) nonnegative matrix. Finally, if $\hat{P} = 0$, then (4.4) implies that E is given by

$$E = -N_{11}P. \tag{4.7}$$

Returning to (3.18), the assumption that F is asymptotically stable implies that $E = \lim_{t\rightarrow\infty} E(t)$ exists for all $E(0)$ and P. Hence we consider the case in which A is also asymptotically stable so that $E = \lim_{t\rightarrow\infty} E(t)$ also exists for $E(t)$ given by (3.19). In this case, the steady-state solution to (3.19) is given by (2.11), that is,

$$E = -A^{-1}P. \tag{4.8}$$

Requiring that the steady-state values given by (4.7) and (4.8) be equal yields

$$A^{-1} = N_{11}. \tag{4.9}$$

The matrix A given by (4.9) will be called the *derived* model. However, for the derived model to exist, N_{11} must be nonsingular. The following result addresses this question.

PROPOSITION 4.1. *N_{11} is nonsingular if and only if M_{22} is nonsingular. In this case*

$$N_{11}^{-1} = M_{11} - M_{12}M_{22}^{-1}M_{21}, \tag{4.10}$$

$$M_{22}^{-1} = N_{22} - N_{21}N_{11}^{-1}N_{12}. \tag{4.11}$$

Proof. The result follows immediately from $M^{-1}M = I_{n^2}$. □

COROLLARY 4.1. *Suppose that M_{22} is nonsingular. Then $M_{12}M_{22}^{-1}M_{21}$ is real.*

Proof. From (4.10) it follows that $M_{12}M_{22}^{-1}M_{21} = M_{11} - N_{11}^{-1}$. Since M_{11} and N_{11}^{-1} are real (in fact, $-N_{11}$ is nonnegative), $M_{12}M_{22}^{-1}M_{21}$ is also real. □

Hence if M_{22} is nonsingular, then (4.9) and (4.10) imply that A is given by

$$A = M_{11} - M_{12}M_{22}^{-1}M_{21}. \tag{4.12}$$

It remains to be shown, however, that A given by (4.12) is asymptotically stable and represents the dynamics of a compartmental model as defined by (2.4), (2.5). For convenience in discussing (4.12) define

$$\mu \triangleq -M_{11}, \qquad \mathscr{P} \triangleq M_{12}M_{22}^{-1}M_{21} \tag{4.13}$$

so that (4.12) can be written as

$$A = -(\mu + \mathscr{P}). \tag{4.14}$$

The key to analyzing (4.14) is to exploit the structure of M_{22}^{-1}. To facilitate our analysis, decompose F as $F = \{F\} + \langle F\rangle$ so that M_{22} can be written as

$$M_{22} = L_d + L_o, \tag{4.15}$$

where

$$L_d \triangleq U_o(\{\bar{F}\} \oplus \{F\})U_o^T, \qquad L_o \triangleq U_o(\langle\bar{F}\rangle \oplus \langle F\rangle)U_o^T.$$

Note that L_d and L_o depend upon the diagonal and off-diagonal portions of F, respectively. When L_d and M_{22} are nonsingular, we use the decomposition (4.15) and consider the identity

$$M_{22}^{-1} = L_d^{-1} - L_d^{-1}L_oM_{22}^{-1}, \tag{4.16}$$

which implies that

$$M_{12}M_{22}^{-1}M_{21} = M_{12}L_d^{-1}M_{21} - M_{12}L_d^{-1}L_oM_{22}^{-1}M_{21}. \tag{4.17}$$

Let us rewrite (4.17) as

$$\mathscr{P} = \mathscr{P}_0 + \mathscr{R}_0, \tag{4.18}$$

where

$$\mathscr{P}_0 \triangleq M_{12}L_d^{-1}M_{21}, \qquad \mathscr{R}_0 \triangleq -M_{12}L_d^{-1}L_oM_{22}^{-1}M_{21}. \tag{4.19}$$

In (4.18), $\mathscr{P}_0$ can be viewed as the zeroth-order term in an expansion of $\mathscr{P}$ while $\mathscr{R}_0$ is the corresponding remainder. To evaluate $\mathscr{P}_0$, define the Hermitian matrix $\Gamma \in \mathbb{C}^{n \times n}$ by

$$\Gamma_{ij} \triangleq \frac{1}{\bar{F}_{ii} + F_{jj}}, \qquad i, j = 1, \dots, n. \tag{4.20}$$

PROPOSITION 4.2. *Suppose that L_d is nonsingular. Then*

$$\mathscr{P}_0 = 2 \operatorname{Re} [\{\langle \Gamma \circ F \rangle F\} + \langle \Gamma \circ F \circ \bar{F} \rangle]. \tag{4.21}$$

Proof. First note that

$$L_d^{-1} = U_o\left[\sum_{i,j=1}^{n} \Gamma_{ij}\mathscr{E}_{ii} \otimes \mathscr{E}_{jj}\right]U_o^T.$$

Then using (A.4), (A.11), (A.13), and $\bar{\Gamma}_{ij} = \Gamma_{ji}$, we have

$$\begin{aligned}
\mathscr{P}_0 &= U_d(\bar{F} \oplus F)\left[\sum_{\substack{i,j=1\\ i \neq j}}^{n} \Gamma_{ij}\mathscr{E}_{ii} \otimes \mathscr{E}_{jj}\right](\bar{F} \oplus F)U_d^T \\
&= \sum_{\substack{i,j=1\\ i \neq j}}^{n} \Gamma_{ij}[(F\mathscr{E}_{jj}F) \circ \mathscr{E}_{ii} + (\bar{F}\mathscr{E}_{ii}\bar{F}) \circ \mathscr{E}_{jj} + (\mathscr{E}_{jj}F) \circ (\bar{F}\mathscr{E}_{ii}) + (\mathscr{E}_{ii}\bar{F}) \circ (F\mathscr{E}_{jj})] \\
&= \sum_{\substack{i,j=1\\ i \neq j}}^{n} \Gamma_{ij}[F_{ij}F_{ji}\mathscr{E}_{ii} + \bar{F}_{ji}\bar{F}_{ij}\mathscr{E}_{jj} + F_{ji}\bar{F}_{ji}\mathscr{E}_{ji} + F_{ij}\bar{F}_{ij}\mathscr{E}_{ij}] \\
&= 2 \operatorname{Re} \sum_{\substack{i,j=1\\ i \neq j}}^{n} [\Gamma_{ij}F_{ij}F_{ji}\mathscr{E}_{ii} + \Gamma_{ij}F_{ij}\bar{F}_{ij}\mathscr{E}_{ij}] \\
&= 2 \operatorname{Re}\left[\sum_{i=1}^{n} \sum_{\substack{j=1\\ j \neq i}}^{n} (\Gamma \circ F)_{ij}F_{ji}\mathscr{E}_{ii} + \sum_{\substack{i,j=1\\ i \neq j}}^{n} (\Gamma \circ F \circ \bar{F})_{ij}\mathscr{E}_{ij}\right] \\
&= 2 \operatorname{Re} [\{\langle \Gamma \circ F \rangle F\} + \langle \Gamma \circ F \circ \bar{F} \rangle]. \qquad \square
\end{aligned}$$

COROLLARY 4.2. *Suppose that $\{F\}$ is asymptotically stable. Then L_d is nonsingular and $\mathscr{P}_0$ is a Z-matrix. If, in addition, F has no zero off-diagonal elements, then $\mathscr{P}_0$ is essentially negative (has negative off-diagonal elements).*

Proof. If $\{F\}$ is asymptotically stable, then L_d is nonsingular. The result now follows from the fact that for $i \neq j$

$$(\mathcal{P}_0)_{ij} = 2|F_{ij}|^2 \operatorname{Re}(F_{ii} + F_{jj})/|\bar{F}_{ii} + F_{jj}|^2 \leqq 0.$$

If, in addition, $F_{ij} \neq 0$, then $(\mathcal{P}_0)_{ij} < 0$. □

Next let us define

$$F_\alpha = \{F\} + \alpha\langle F\rangle, \tag{4.22}$$

where $\{F\}$ and $\alpha\langle F\rangle$ are diagonal and off-diagonal matrices, respectively, and α is a positive number. The scalar α in (4.22) allows us to adjust the strength of the off-diagonal coupling in F_α. When F is replaced by F_α, derived quantities such as M_{22} and $\mathcal{P}_0$ are written as $M_{22\alpha}$ and $\mathcal{P}_{0\alpha}$. Consequently, (4.12) becomes

$$A_\alpha = M_{11} - M_{12\alpha}M_{22\alpha}^{-1}M_{21\alpha}. \tag{4.23}$$

We thus have the following result for the case of weak coupling, that is, for $\alpha \approx 0$.

COROLLARY 4.3. *Suppose that $\{F\}$ is asymptotically stable and F has no zero off-diagonal elements. Then there exists $\alpha_0 > 0$ such that F_α is asymptotically stable, $M_{22\alpha}$ is nonsingular, and $\mathcal{P}_\alpha$ is essentially negative for all $\alpha \in (0, \alpha_0)$.*

Proof. By Corollary 4.2, $\mathcal{P}_{0\alpha} = M_{12\alpha}L_d^{-1}M_{21\alpha}$ is essentially negative for all $\alpha > 0$. Furthermore, it follows from (3.10) and (3.11) that $\mathcal{P}_{0\alpha} = O(\alpha^2)$ as $\alpha \to 0$. Thus if α is sufficiently small, then F_α is asymptotically stable and $M_{22\alpha} = L_d + \alpha L_o$ is nonsingular. Consequently, $\mathcal{R}_{0\alpha} = O(\alpha^3)$. The result now follows from the fact that $\mathcal{P}_\alpha = \mathcal{P}_{0\alpha} + \mathcal{R}_{0\alpha}$. □

THEOREM 4.1. *Suppose that $\{F\}$ is asymptotically stable and F has no zero off-diagonal elements. Then there exists $\alpha_0 > 0$ such that $-A_\alpha$ is a nonsingular M-matrix for all $\alpha \in [0, \alpha_0)$.*

Proof. Let $\alpha \in [0, \alpha_0)$, where α_0 is given by Corollary 4.3. Since by (4.14) $-\langle A_\alpha\rangle = \langle \mathcal{P}_\alpha\rangle$, it follows from Corollary 4.3 that $-A_\alpha$ is a Z-matrix. Furthermore, by (4.6) and (4.9), $-A_\alpha^{-1} = \int_0^\infty e^{F_\alpha t} \circ e^{\bar{F}_\alpha t}\, dt$ is nonnegative. Hence it follows from [6, p. 137], that $-A_\alpha$ is a nonsingular M-matrix. □

COROLLARY 4.4. *Under the assumptions of Theorem 4.1, A_α is asymptotically stable for all $\alpha \in [0, \alpha_0)$.*

Proof. By [6, p. 135], each eigenvalue of $-A_\alpha$ has a positive real part. Hence A_α is asymptotically stable. □

The remainder of this section is devoted to further analysis of the properties of $\mathcal{P}$. The following result gives an alternative sufficient condition for M_{22} to be asymptotically stable and hence nonsingular.

PROPOSITION 4.3. *If F is upper triangular, then M_{22} is asymptotically stable.*

Proof. The result follows from the fact that L_d is asymptotically stable and L_o is strictly upper triangular. □

PROPOSITION 4.4. *Suppose that M_{22} is nonsingular. If F is symmetric (but not necessarily either real or Hermitian), then $\mathcal{P}$ is symmetric. If, in addition, $\mathcal{P}$ is a Z-matrix, then A defined by* (4.14) *is negative definite.*

Proof. If F is symmetric, then so are $\langle F\rangle$ and $\langle \bar{F}\rangle$. The result is now immediate. □

Next, we consider the case in which $\langle F\rangle$ is skew-Hermitian. This case arises frequently in applications in which the modal coupling is energy conservative.

PROPOSITION 4.5. *Suppose that M_{22} is nonsingular. If $\langle F\rangle$ is skew-Hermitian, then*

$$\mathscr{P}\mathbf{e} = \mathscr{P}^T\mathbf{e} = 0. \tag{4.24}$$

If, in addition, $\operatorname{Re}\langle F\rangle = 0$, *then $\mathscr{P}$ is symmetric.*

Proof. Using the fact that $\operatorname{vecd} I_n = \mathbf{e}$ along with (A.14) and (A.16) yields

$$\begin{aligned}(\langle \bar{F}\rangle \oplus \langle F\rangle)U_d^T\mathbf{e} &= (\langle \bar{F}\rangle \oplus \langle F\rangle)U_d^T \operatorname{vecd} I_n\\ &= (\langle \bar{F}\rangle \oplus \langle F\rangle)\operatorname{vec} I_n\\ &= \operatorname{vec}(\langle F\rangle + \langle \bar{F}\rangle^T)\\ &= 0,\end{aligned}$$

which along with (3.11) and (4.13) implies that $\mathscr{P}\mathbf{e} = 0$. A similar argument shows that $\mathscr{P}^T\mathbf{e} = 0$. If $\operatorname{Re}\langle F\rangle = 0$, then $\langle F\rangle = j\hat{F}$, where $\hat{F}$ is real. Since $\langle F\rangle$ is skew-Hermitian, it follows that $\hat{F}$ is symmetric. Consequently, $\langle F\rangle$ is symmetric, which implies that F is symmetric. Now by Proposition 4.4, $\mathscr{P}$ is symmetric. □

5. Compartmental modeling of state space systems. In this section we relate the steady-state second-moment analysis of § 4 to the compartmental model discussed in § 2.

If $-A$ is a nonsingular M-matrix (assuming M_{22} is nonsingular), it follows from property (M_{36}) [6, p. 137], that there exists a diagonal matrix $D = \operatorname{diag}(d_1, \ldots, d_n)$ with positive diagonal elements $d_i > 0$, $i = 1, \ldots, n$, such that $D(-A)D^{-1}$ is strictly diagonally dominant, that is,

$$-A_{ii} > \sum_{\substack{j=1\\ j\neq i}}^{n} (d_j/d_i)A_{ji}. \tag{5.1}$$

Note that $-A_{ii}$ is positive since a nonsingular M-matrix has positive diagonal elements. Now define $\hat{A} \triangleq DAD^{-1}$ and note that $-\hat{A}$ is also a nonsingular M-matrix. To show that $\hat{A}$ has the form of a compartmental model, define

$$\sigma_{ii} \triangleq -A_{ii} - \sum_{\substack{j=1\\ j\neq i}}^{n} (d_j/d_i)A_{ji}, \qquad i = 1, \ldots, n, \tag{5.2}$$

$$\sigma_{ij} \triangleq (d_i/d_j)A_{ij}, \quad i \neq j, \quad i, j = 1, \ldots, n. \tag{5.3}$$

With (5.3) we can rewrite (5.2) as

$$\sigma_{ii} = -A_{ii} - \sum_{\substack{j=1\\ j\neq i}}^{n} \sigma_{ji}. \tag{5.4}$$

Since $\hat{A}_{ij} = (d_i/d_j)A_{ij}$, it follows that

$$\hat{A}_{ii} = \sum_{j=1}^{n} \sigma_{ji}, \qquad i = 1, \ldots, n, \tag{5.5}$$

$$\hat{A}_{ij} = \sigma_{ij}, \quad i \neq j, \quad i, j = 1, \ldots, n, \tag{5.6}$$

which verifies (2.4), (2.5) with A replaced by $\hat{A}$.

Next, we introduce the scaled energy states $E_s \triangleq DE$ and scaled power input $P_s \triangleq DP$ so that (4.8) becomes

$$0 = \hat{A}E_s + P_s. \tag{5.7}$$

With this scaling and the definition of $\hat{A}$, (5.7) has the form of a steady-state compartmental model as given by (2.12).

Remark 5.1. Consider the energy conservative case in which $\langle F \rangle$ is skew-Hermitian so that (4.23) holds. Then the row and column sums of $\mathscr{P}$ are zero. If $\mathscr{P}$ is also a Z-matrix, then no scaling is required (that is, $D = I$ suffices) to obtain a steady-state compartmental model.

6. Conclusion. Compartmental models are widely used to represent the dynamics of systems involving the exchange of inherently nonnegative quantities such as mass or energy. In this paper we summarized some of the key properties of these models and characterized their steady-state energy distribution. In addition, conditions were given under which the steady-state energy distribution tends toward equipartition in the limit of strong off-diagonal coupling. We then considered arbitrary state space models with additive white noise disturbance and obtained an equation that governs the evolution of the nonnegative diagonal system. The steady-state limit of this diagonal system was then examined for its relationship to steady-state compartmental models. The key step in this regard was to show that the coefficient matrix is a Z-matrix, that is, has nonpositive off-diagonal elements. It was shown that if the off-diagonal coupling is sufficiently weak, then (up to a positive scaling) the diagonal system does in fact have the form of a compartmental model. Conditions under which the diagonal system is a compartmental model in the case of strong coupling remain an area for future research.

Appendix. Kronecker matrix algebra. In this Appendix we review some basic definitions and identities from Kronecker matrix algebra. Our main references are [67] and [68]. We also introduce several specialized definitions that are specific to this paper.

For $A \in \mathbb{C}^{n\times m}$, let $\operatorname{col}_i(A)$ denote the ith column of A and define the vec and vec^{-1} operators by

$$\operatorname{vec} A \triangleq \begin{bmatrix} \operatorname{col}_1(A) \\ \vdots \\ \operatorname{col}_m(A) \end{bmatrix} \in \mathbb{C}^{nm}, \qquad \operatorname{vec}^{-1}(\operatorname{vec} A) \triangleq A.$$

For $A \in \mathbb{C}^{n\times m}$ and $B \in \mathbb{C}^{p\times q}$, the Kronecker product of A and B is defined by

$$A \otimes B \triangleq \begin{bmatrix} A_{11}B & A_{12}B & \cdots & A_{1m}B \\ A_{21}B & A_{22}B & \cdots & A_{2m}B \\ \vdots & \vdots & & \vdots \\ A_{n1}B & A_{n2}B & \cdots & A_{nm}B \end{bmatrix} \in \mathbb{R}^{np\times mq},$$

while for $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{m\times m}$ the Kronecker sum of A and B is defined by

$$A \oplus B \triangleq A \oplus I_m + I_n \oplus B \in \mathbb{C}^{nm\times nm}.$$

For compatible complex matrices A, B, C, D, the following identities hold:

$$(A + B) \otimes C = A \otimes C + B \otimes C, \tag{A.1}$$

$$A \otimes (B + C) = A \otimes B + A \otimes C, \tag{A.2}$$

$$(A \otimes B)^T = A^T \otimes B^T, (A \oplus B)^T = A^T \oplus B^T, \tag{A.3}$$

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD), \tag{A.4}$$

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}, \tag{A.5}$$

$$\operatorname{vec} ABC = (C^T \otimes A) \operatorname{vec} B, \tag{A.6}$$

$$\operatorname{vec} (AB + BC) = (C^T \oplus A) \operatorname{vec} B, \tag{A.7}$$

$$e^{A \oplus B} = e^A \otimes e^B, \tag{A.8}$$

$$(A \otimes B) \circ (C \otimes D) = (A \circ C) \otimes (B \circ D). \tag{A.9}$$

If $w, y \in \mathbb{C}^n$ and $x, z \in \mathbb{C}^m$, then

$$(wx^T) \circ (yz^T) = (w \circ y)(x \circ z)^T. \tag{A.10}$$

Next, let e_i denote the ith column of the $n \times n$ identity matrix, where the dimension n is determined by context, and define $\mathscr{E}_{rs} \triangleq e_r e_s^T$, which is the not-necessarily square matrix whose (r, s) element is 1 and whose remaining elements are all 0. Now define the $n \times n^2$ matrix

$$U_d \triangleq [\mathscr{E}_{11} \mathscr{E}_{22} \cdots \mathscr{E}_{nn}] = \sum_{i=1}^{n} e_i^T \otimes \mathscr{E}_{ii}.$$

Furthermore, letting $I_{/i}$ denote the $(n-1) \times n$ matrix obtained by deleting the ith row of the $n \times n$ identity matrix, define the $(n^2 - n) \times n^2$ matrix

$$U_o \triangleq \begin{bmatrix} I_{/1} & & & \\ & I_{/2} & 0 & \\ & 0 & \ddots & \\ & & & I_{/n} \end{bmatrix} = \sum_{i=1}^{n} \mathscr{E}_{ii} \otimes I_{/i}.$$

If $A, B \in \mathbb{C}^{n \times n}$, then [69] and [70]

$$U_d(A \otimes B) U_d^T = U_d(B \otimes A) U_d^T = A \circ B. \tag{A.11}$$

Note that U_d and U_o satisfy the identities

$$U_d U_d^T = I_n, \quad U_d U_o^T = 0_{n \times (n^2 - n)}, \quad U_o U_o^T = I_{n^2 - n}, \tag{A.12}$$

$$U_o^T U_o = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \mathscr{E}_{ii} \otimes \mathscr{E}_{jj}. \tag{A.13}$$

Hence the matrix $U \in \mathbb{R}^{n^2 \times n^2}$ defined by

$$U \triangleq \begin{bmatrix} U_d \\ U_o \end{bmatrix}$$

satisfies $U^T = U^{-1}$; that is, U is an orthogonal (but not symmetric) permutation matrix.

For a square matrix $A \in \mathbb{C}^{n \times n}$, let $\{A\}$ and $\langle A \rangle$ denote the diagonal and off-diagonal portions of A, respectively. That is, $\{A\}$ is the diagonal matrix defined by

$$\{A\} \triangleq I \circ A,$$

and $\langle A \rangle$ is the off-diagonal matrix defined by

$$\langle A \rangle \triangleq A - \{A\}.$$

Next, as in [67], let vecd A denote the n-vector comprised of the diagonal elements of A, that is,

$$\operatorname{vecd} A \triangleq \begin{bmatrix} A_{11} \\ \vdots \\ A_{nn} \end{bmatrix} = \{A\}\mathbf{e}, \qquad \operatorname{vecd}^{-1}(\operatorname{vecd} A) \triangleq \{A\},$$

where $\mathbf{e} \triangleq [1\ 1\ \cdots\ 1]^T$, and let veco A denote the $(n^2 - n)$-vector comprised of the off-diagonal elements of A ordered in accordance with vec A. Define also $\operatorname{veco}^{-1}(\operatorname{veco} A) \triangleq \langle A \rangle$. The above-defined operators satisfy the following identities:

(A.14) $$\operatorname{vecd} A = U_d \operatorname{vec} A = \operatorname{vecd}\{A\} = U_d \operatorname{vec}\{A\},$$

(A.15) $$\operatorname{veco} A = U_o \operatorname{vec} A = \operatorname{veco}\langle A\rangle = U_o \operatorname{vec}\langle A\rangle,$$

(A.16) $$\operatorname{vec}\{A\} = U_d^T \operatorname{vecd} A, \qquad \operatorname{vec}\langle A\rangle = U_o^T \operatorname{veco} A.$$

Finally, if $A, B \in \mathbb{C}^{n \times n}$, it is useful to note that

(A.17) $$\{A \oplus B\} = \{A\} \oplus \{B\}, \qquad \langle A \oplus B\rangle = \langle A\rangle \oplus \langle B\rangle,$$

(A.18) $$A \oplus B = \{A\} \oplus \{B\} + \langle A\rangle \oplus \langle B\rangle,$$

(A.19) $$U_d(A \oplus B)U_d^T = \{A + B\}, \qquad U_o\{A \oplus B\}U_d^T = 0,$$

(A.20) $$U_o(A \oplus B)U_d^T = U_o(\langle A\rangle \oplus \langle B\rangle)U_d^T.$$

Acknowledgment. We wish to thank Y. Kishimoto for several helpful suggestions.

REFERENCES

[1] J. Z. HEARON, *Theorems on linear systems*, Ann. New York Acad. Sci., 108 (1963), pp. 36–67.

[2] R. R. MOHLER, *Biological modeling with variable compartmental structure*, IEEE Trans. Automat. Control, 19 (1974), pp. 922–926.

[3] H. MAEDA, S. KODAMA, AND F. KAJIYA, *Compartmental system analysis: Realization of a class of linear systems with physical constraints*, IEEE Trans. Circuits and Systems, 24 (1977), pp. 8–14.

[4] I. W. SANDBERG, *On the mathematical foundations of compartmental analysis in biology, medicine, and ecology*, IEEE Trans. Circuits and Systems, 25 (1978), pp. 273–279.

[5] H. MAEDA, S. KODAMA, AND Y. OHTA, *Asymptotic behavior of nonlinear compartmental systems: Nonoscillation and stability*, IEEE Trans. Circuits and Systems, 25 (1978), pp. 372–378.

[6] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[7] R. E. FUNDERLIC AND J. B. MANKIN, *Solution of homogeneous systems of linear equations arising from compartmental models*, SIAM J. Sci. Statist. Comp., 2 (1981), pp. 375–383.

[8] J. W. NIEUWENHUIS, *About nonnegative realizations*, Systems Control Lett., 1 (1982), pp. 283–287.

[9] D. H. ANDERSON, *Compartmental Modeling and Tracer Kinetics*, Springer-Verlag, Berlin, New York, 1983.

[10] K. GODFREY, *Compartmental Models and Their Applications*, Academic Press, New York, 1983.

[11] Y. OHTA, H. MAEDA, AND S. KODAMA, *Reachability, observability, and realizability of continuous-time positive systems*, SIAM J. Control Optim., 22 (1984), pp. 171–180.

[12] J. A. JACQUEZ, *Compartmental Analysis in Biology and Medicine*, 2nd ed., University of Michigan Press, Ann Arbor, 1985.

[13] A. BERMAN, M. NEUMANN, AND R. J. STERN, *Nonnegative Matrices in Dynamic Systems*, Wiley & Sons, New York, 1989.

[14] J. A. MORRISON AND J. MCKENNA, *Coupled line equations with random coupling*, Bell Systems Tech. J., 51 (1972), pp. 209–228.

[15] R. BURRIDGE AND G. C. PAPANICOLAOU, *The geometry of coupled mode propagation in one-dimensional random media*, Comm. Pure Appl. Math., 25 (1972), pp. 715–757.

[16] G. C. Papanicolaou, *A kinetic theory for power transfer in stochastic systems*, J. Math. Phys., 13 (1972), pp. 1912–1918.

[17] W. Kohler and G. C. Papanicolaou, *Power statistics for wave propagation in one dimension and comparison with radiative transport theory*, J. Math. Phys., 14 (1973), pp. 1733–1745.

[18] J. C. Willems, *Dissipative dynamical systems part* I: *General theory*, Arch. Rational Mech. Anal., 45 (1972), pp. 321–351.

[19] J. C. Willems, *Dissipative dynamical systems part* II: *Linear systems with quadratic supply rates*, Arch. Rational Mech. Anal., 45 (1972), pp. 352–393.

[20] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis*: *A Modern Systems Theory Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[21] R. W. Brockett and J. C. Willems, *Stochastic control and the second law of thermodynamics*, Proc. Conf. Dec. Contr., San Diego, CA, 1978, pp. 1007–1011.

[22] D. J. Hill and P. J. Moylan, *Dissipative dynamical systems*: *Basic input-output and state properties*, J. Franklin Inst., 309 (1980), pp. 327–357.

[23] B. D. O. Anderson, *Nonlinear networks and Onsager-Casimir reversibility*, IEEE Trans. Circuits and Systems, 27 (1980), pp. 1051–1058.

[24] J. L. Wyatt, Jr., L. O. Chua, J. W. Gannett, I. C. Göknar, and D. N. Green, *Energy concepts in the state-space theory of nonlinear n-ports*: *Part* I–*Passivity*, IEEE Trans. Circuits and Systems, 28 (1981), pp. 48–61.

[25] ———, *Energy concepts in the state-space theory of nonlinear n-ports*: *Part* II–*Losslessness*, IEEE Trans. Circuits and Systems, 29 (1982), pp. 417–430.

[26] J. L. Wyatt, Jr., W. M. Siebert, and J.-N. Tan, *A frequency domain inequality for stochastic power flow in linear networks*, IEEE Trans. Circuits and Systems, 31 (1984), pp. 809–814.

[27] M. S. Gupta, *Upper bound on the rate of entropy increase accompanying noise power flow through linear systems*, IEEE Trans. Circuits and Systems, 35 (1988), pp. 1162–1163.

[28] D. Karnopp, *Power-conserving transformations*: *Physical interpretations and applications using bond graphs*, J. Franklin Inst., 288 (1969), pp. 175–201.

[29] D. Karnopp, *The energetic structure of multibody dynamic systems*, J. Franklin Inst., 306 (1978), pp. 165–181.

[30] R. C. Rosenberg and D. C. Karnopp, *Introduction to Physical System Dynamics*, McGraw-Hill, New York, 1983.

[31] R. H. Lyon and G. Maidanik, *Power flow between linearly coupled oscillators*, J. Acoust. Soc. Amer., 34 (1962), pp. 623–639.

[32] E. E. Ungar, *Statistical energy analysis of vibrating systems*, Trans. ASME, 89 (1967), pp. 626–632.

[33] T. D. Scharton and R. H. Lyon, *Power flow and energy sharing in random vibration*, J. Acoust. Soc. Amer., 34 (1968), pp. 1332–1343.

[34] R. H. Lyon, *Statistical Energy Analysis of Dynamical Systems*: *Theory and Applications*, MIT Press, Cambridge, MA, 1975.

[35] P. W. Smith, Jr., *Statistical models of coupled dynamical systems and the transition from weak to strong coupling*, J. Acoust. Soc. Amer., 65 (1979), pp. 695–698.

[36] J. Woodhouse, *An approach to the theoretical background of statistical energy analysis applied to structural vibration*, J. Acoust. Soc. Amer., 69 (1981), pp. 1695–1709.

[37] F. J. Fahy, *Statistical energy analysis*, in Noise and Vibration, R. White and J. Walker, eds., Ellis Horwood Ltd., Chichester, 1982.

[38] E. H. Dowell and Y. Kubota, *Asymptotic modal analysis and statistical energy analysis of dynamical systems*, J. Appl. Mech., 52 (1985), pp. 949–957.

[39] C. H. Hodges and J. Woodhouse, *Theories of noise and vibration transmission in complex structures*, Rep. Progr. Phys., 49 (1986), pp. 107–170.

[40] A. J. Keane and W. G. Price, *Statistical energy analysis of strongly coupled systems*, J. Sound Vibration, 117 (1987), pp. 363–386.

[41] R. K. Pearson and T. L. Johnson, *Energy equipartition and fluctuation-dissipation theorems for damped flexible structures*, Quart. Appl. Math., 45 (1987), pp. 223–238.

[42] R. S. Langley, *A general derivation of the statistical energy analysis equations for coupled dynamic systems*, J. Sound Vibration, 135 (1989), pp. 499–508.

[43] D. W. Miller and A. H. von Flotow, *A travelling wave approach to power flow in structural networks*, J. Sound Vibration, 128 (1989), pp. 145–162.

[44] I. J. Busch-Vishniac and H. M. Paynter, *Bond graph models of sound and vibration systems*, J. Acoust. Soc. Amer., 85 (1989), pp. 1750–1758.

[45] D. G. MacMartin and S. R. Hall, *An H_∞ power flow approach to the control of uncertain structures*, Proc. Amer. Control Conf., San Diego, CA, 1990, pp. 3073–3080.

[46] D. C. HYLAND AND D. S. BERNSTEIN, *Power flow, energy balance, and statistical energy analysis for large scale, interconnected systems*, Proc. Amer. Control Conf., San Diego, CA, 1990, pp. 1929–1934.

[47] V. E. BENES, *A 'thermodynamic' theory of traffic in connecting networks*, Bell Systems Tech. J., 42 (1963), pp. 567–607.

[48] M. SCHWARTZ, *Telecommunication Networks: Protocols, Modeling and Analysis*, Addison-Wesley, Reading, MA, 1989.

[49] N. W. HAGOOD AND E. P. CRAWLEY, *Cost Averaging Techniques for Robust Control of Parametrically Uncertain Systems*, SERC Report # 9-91, MIT, Cambridge, MA, June 1991.

[50] S. R. HALL, D. G. MACMARTIN, AND D. S. BERNSTEIN, *Covariance averaging in the analysis of uncertain systems*, Proc. Conf. on Decision and Control, Tucson, AZ, December 1992, pp. 1842–1859.

[51] W. E. HOPKINS, JR., *Optimal control of linear systems with parameter uncertainty*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 72–74.

[52] G. W. STEWART, *Stochastic perturbation theory*, SIAM Rev., 32 (1990), pp. 579–610.

[53] E. C. HOWE AND C. R. JOHNSON, *Expected-value norms on matrices*, Linear Algebra Appl., 139 (1990), pp. 21–29.

[54] A. N. MICHEL AND R. K. MILLER, *Qualitative Analysis of Large Scale Dynamical Systems*, Academic Press, New York, 1977.

[55] D. D. SILJAK, *Large-Scale Dynamic Systems*, North-Holland, Amsterdam, 1978.

[56] M. VIDYASAGAR, *Input-Output Analysis of Large-Scale Interconnected Systems*, Springer-Verlag, Berlin, New York, 1981.

[57] D. D. SILJAK, *Complex dynamical systems: Dimensionality, structure, and uncertainty*, Large Scale Systems, 4 (1983), pp. 279–294.

[58] L. T. GRUJIC, A. A. MARTYNYUK, AND M. RIBBONS-PARELLA, *Large Scale Systems Stability Under Structural and Singular Perturbations*, Springer-Verlag, Berlin, New York, 1987.

[59] J. LUNZE, *Stability analysis of large-scale systems composed of strongly coupled similar subsystems*, Automatica, 25 (1989), pp. 561–570.

[60] G. DAHLQUIST, *On matrix majorants and minorants, with application to differential equations*, Linear Algebra Appl., 52/53 (1983), pp. 199–216.

[61] S. L. CAMPBELL AND C. D. MEYER, JR., *Generalized Inverses of Linear Transformations*, Pitman, Boston, 1979.

[62] A. BERMAN, R. S. VARGA, AND R. C. WARD, *ALPS: Matrices with nonpositive off-diagonal entries*, Linear Algebra Appl., 21 (1978), pp. 233–244.

[63] C. D. MEYER, JR. AND M. W. STADELMAIER, *Singular M-matrices and inverse positivity*, Linear Algebra Appl., 22 (1978), pp. 139–156.

[64] J. R. MAGNUS AND H. NEUDECKER, *Matrix Differential Calculus*, Wiley & Sons, New York, 1988.

[65] B. F. LAMOND AND M. L. PUTERMAN, *Generalized inverses in discrete time Markov decision processes*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 118–134.

[66] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley, New York, 1972.

[67] J. W. BREWER, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits and Systems, 25 (1978), pp. 772–781.

[68] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, Wiley, New York, 1981.

[69] M. MARCUS AND N. KHAN, *A note on the Hadamard product*, Canad. Math. Bull., 2 (1959), pp. 81–83.

[70] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 903–926, October 1993

DERIVATIVES OF EIGENVALUES AND EIGENVECTORS OF MATRIX FUNCTIONS*

ALAN L. ANDREW[†], K.-W. ERIC CHU[‡], AND PETER LANCASTER[§]

Abstract. For an $n \times n$ matrix-valued function $L(\boldsymbol{\rho}, \lambda)$, where $\boldsymbol{\rho}$ is a vector of independent parameters and λ is an eigenparameter, the eigenvalue-eigenvector problem has the form $L(\boldsymbol{\rho}, \lambda(\boldsymbol{\rho}))\mathbf{x}(\boldsymbol{\rho}) = \mathbf{0}$. Real or complex values for $\boldsymbol{\rho}$ and λ are admitted, and L is assumed to depend analytically on these variables. In particular, nonlinear dependence on λ is the main concern. On the assumption that the eigenvalue-eigenvector problem itself can be satisfactorily solved, a study is made of the derivatives (sensitivities) of λ and $\mathbf{x}$ with respect to $\boldsymbol{\rho}$. Analysis is made of the existence of derivatives, the effect of normalization strategies, and the solvability and condition of the bordered matrix equations arising naturally in this context. Implications for various classical eigenvalue problems $L(\boldsymbol{\rho}, \lambda) = A(\boldsymbol{\rho}) - \lambda I$ are clarified.

Key words. nonlinear eigenvalue problems, eigenvalue and eigenvector sensitivities, generalized inverses

AMS subject classifications. 65F15, 15A18

1. Introduction and hypotheses. In this paper we study $n \times n$ matrix-valued functions $L(\rho_1, \dots, \rho_m, \lambda)$ where $\rho_1, \dots, \rho_m$ are either real or complex independent variables, and $n \geq 2$ and λ is an eigenvalue parameter that may also be real or complex valued. For brevity, let $\boldsymbol{\rho}$ denote the m-tuple $\rho_1, \dots, \rho_m$ and write $L(\rho_1, \dots, \rho_m, \lambda) = L(\boldsymbol{\rho}, \lambda)$.

The number $\lambda(\boldsymbol{\rho})$ is an *eigenvalue* of L at $\boldsymbol{\rho}$ if the matrix $L(\boldsymbol{\rho}, \lambda(\boldsymbol{\rho}))$ is singular. Thus, eigenvalues have associated *eigenvectors* $\mathbf{x}(\boldsymbol{\rho}) \neq \mathbf{0}$ such that

$$L(\boldsymbol{\rho}, \lambda(\boldsymbol{\rho}))\mathbf{x}(\boldsymbol{\rho}) = \mathbf{0}. \tag{1.1}$$

We may also call $\mathbf{x}(\boldsymbol{\rho})$ a *right eigenvector* of L at $\boldsymbol{\rho}$. A *left eigenvector* of L at $\boldsymbol{\rho}$ is a right eigenvector of the transpose L^T.

Let $\mathcal{D}$ be a domain in a space of k variables $z_1, \dots, z_k$. A real or complex valued function $f(\mathbf{z}), \mathbf{z} = (z_1, \dots, z_k)$, defined on $\mathcal{D}$, is said to be an *analytic function of* $\mathbf{z}$ *on* $\mathcal{D}$ if, to each $\mathbf{z}_0 = (z_1^0, \dots, z_k^0) \in \mathcal{D}$, there corresponds a neighbourhood $\mathcal{N}(\mathbf{z}_0) \subset \mathcal{D}$ in which $f(\mathbf{z})$ is the sum of an absolutely convergent power series in $(z_1 - z_1^0), \dots, (z_k - z_k^0)$. Here, and in the sequel, we follow Bochner and Martin [BM].

Our fundamental hypotheses on the function L may now be formulated as follows: The variables $\boldsymbol{\rho}, \lambda$ belong to domains $\mathcal{N}, \mathcal{N}_\lambda$, respectively, such that:

(i) The elements of L are analytic functions of the $m+1$ variables $\rho_1, \dots, \rho_m, \lambda$ on $\mathcal{N} \times \mathcal{N}_\lambda$.

*Received by the editors April 1, 1991; accepted for publication (in revised form) December 30, 1991.

[†]Department of Mathematics, La Trobe University, Bundoora, Victoria 3083, Australia (matala@lure.latrobe.edu.au).

[‡]Department of Mathematics, Monash University, Clayton, Victoria 3168, Australia (apm543z@vaxc.cc.monash.edu.au).

[§]Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada, T2N 1N4 (lancaste@acs.ucalgary.ca). The research of this author was supported in part by a grant from the Natural Sciences and Engineeering Research Council of Canada, by a visiting fellowship at Monash University, and by a grant from La Trobe University.

(ii) For each $\boldsymbol{\rho} \in \mathcal{N}$ there is a $\lambda \in \mathcal{N}_\lambda$ such that $\det L(\boldsymbol{\rho}, \lambda) \neq 0$.

Note that, for an $n \times n$ matrix function $A(\boldsymbol{\rho})$, our formulation includes the classical eigenvalue problem if we define $L(\boldsymbol{\rho}, \lambda) = A(\boldsymbol{\rho}) - \lambda I$. Similarly, the "generalized" eigenvalue problem $A\mathbf{x} = \lambda B\mathbf{x}$ is included by defining $L(\boldsymbol{\rho}, \lambda) = A(\boldsymbol{\rho}) - \lambda B(\boldsymbol{\rho})$ and, for parameter-dependent quadratic eigenvalue problems, which arise in many applications ([BL], [CMa], [HA], [L], [Ro]),

$$L(\boldsymbol{\rho}, \lambda) = A(\boldsymbol{\rho})\lambda^2 + B(\boldsymbol{\rho})\lambda + C(\boldsymbol{\rho}). \tag{1.2}$$

Concerning hypothesis (i), note that this degree of generality admits a theory of eigenvalue multiplicities and chains of generalized eigenvectors. This theory is developed in [GR]. Only its most fundamental ideas are used in this work, but we would emphasise that, although the classical eigenvalue problem is an important special case, our major concern is with nonlinear dependence on λ.

Our objective is to study the derivatives of eigenvalues and eigenvectors with respect to the parameters $\rho_1, \ldots, \rho_m$, including some discussion of the problem of existence of these derivatives. The derivatives tell us about the sensitivities of eigenvalues and eigenvectors and are, therefore, of great interest in many applications. There is, of course, a very extensive literature including books and papers addressed to mathematicians and a broad spectrum of scientists and engineers. Much of this is directed to special classes of eigenvalue problems such as those described above and, particularly, to the case of one parameter, $m = 1$. We do not attempt to make a representative list of references here, but let them arise naturally in the body of the paper.

Careful investigations with $m > 1$ are relatively few in number and we believe that none have the degree of generality admitted by hypotheses (i) and (ii). We note the interesting recent work of Sun (see [S2] and references therein). He admits multiple real parameters in his analysis of matrix functions $L(\boldsymbol{\rho}, \lambda) = A(\boldsymbol{\rho}) - \lambda I$ and $L(\boldsymbol{\rho}, \lambda) = A(\boldsymbol{\rho}) - \lambda B(\boldsymbol{\rho})$ and, in contrast to our investigation of analytic eigenfunctions $\lambda(\boldsymbol{\rho})$, he considers the more general question of existence of one-sided directional derivatives for $\lambda(\boldsymbol{\rho})$ (see also [HC]). In §§5 and 6, we also consider the case in which $\lambda(\boldsymbol{\rho})$ and $\mathbf{x}(\boldsymbol{\rho})$ are required to have only partial derivatives with respect to the components of $\boldsymbol{\rho}$.

Because of the linear dependence on $\mathbf{x}$ the problem (1.1) might be described as "mildly" nonlinear. So there is also a strong connection with the large body of literature on nonlinear analysis and, particularly, the literature on continuation methods. (See [Kel], for example.) In the context of the problem (1.1) continuation is also an important concern, and calculation of derivatives can be seen as a first step in that direction. That line of investigation is not pursued here, although some connections are clarified in an appendix.

The contributions of this paper concern mainly nonlinear dependence on λ and include:

(a) A study of the effect, and relative advantages, of different normalization strategies for the eigenvectors.

(b) Explicit formulae for derivatives and their connection with item (a).

(c) Discussion of the solvability and condition of the equations with bordered coefficient matrices, which arise naturally in this context.

These contributions provide insights and guidance relevant to the design of algorithms for the solution of parameter-dependent eigenvalue problems of the form (1.1).

Some sufficient conditions for existence of analytic eigenvalues and eigenvectors are established in §2, more detail on semisimple eigenvalues is provided in §3, while §4 includes a discussion of various normalization strategies (a subject on which there has been some confusion in the engineering literature). The computation of partial derivatives of eigenvalues and eigenvectors is investigated in §§5 and 6, using equations that are no more difficult to solve than those arising in the linear case. This contrasts with the initial computation of eigenvalues and eigenvectors, which is significantly more difficult in the nonlinear case. Sections 7 and 8 establish and examine new explicit formulae for the partial derivatives in terms of the Moore–Penrose inverse and the group inverse, respectively, while §9 extends this work to higher derivatives. Section 10 contains further discussion of the condition of bordered matrices appearing in this work. The extent to which our results simplify in the case of hermitian matrix functions and in the case of the classical linear eigenvalue problem is considered in §§11 and 12, respectively. Some of the results proved here were announced by us without proof in [ACL].

Our results in §8 generalize some results of [MS], which considers simple eigenvalues in the linear case. Our equation (5.5) (but not our analysis) is classical for simple eigenvalues in the linear case and has also been derived in the case of the quadratic eigenvalue problem (again with simple eigenvalues) in [CMa]. Our relaxation of the restriction to simple eigenvalues, though not complete, is important, for example, in transient optimization problems, which are often characterized by root coalescence [Ga], [HC]. If, as is frequently the case, there is no simultaneous coalescence of eigenvectors, then the semisimple hypothesis will be satisfied. Indeed, multiple eigenvalues, once neglected in the engineering literature because of their difficulty, are now receiving increasing attention (see [HA], [HR], [Oj] and the references therein). Our results also cover problems with transcendental dependence on the eigenvalue such as occur, for example, in certain methods for the numerical solution of differential equations (see [A], [OM] and the references therein).

2. Existence of analytic eigenvalues and eigenvectors. Let $f(\rho,\lambda) = \det L(\rho,\lambda)$ where $L(\rho,\lambda)$ satisfies conditions (i) and (ii) of §1. These conditions imply that, for any fixed $\rho_0 \in \mathcal{N}$, the eigenvalues $\lambda(\rho_0)$ are zeros of $f(\rho_0,\lambda)$ (if any) and form a finite or countable set of points in $\mathcal{N}_\lambda$. If the multiplicity of $\lambda(\rho_0)$ as a zero of $f(\rho_0,\lambda)$ is positive and equal to the dimension of $\ker L(\rho_0,\lambda(\rho_0))$, then $\lambda(\rho_0)$ is said to be a *semisimple* eigenvalue.

When $\lambda(\rho_0)$ is a simple zero of $f(\rho_0,\lambda)$, it is known as a *simple* eigenvalue (and is necessarily semisimple). When this is the case, the implicit function theorem can be applied to f at $(\rho_0,\lambda(\rho_0))$ (see [BM] and [C1]), and we see immediately that λ is an analytic function, $\lambda(\rho)$, in some neighbourhood of ρ_0. Consequently, there is a neighbourhood $\mathcal{N}_0$ of ρ_0 in which $L(\rho,\lambda(\rho))$ is an analytic function with rank identically equal to $n-1$.

Now recall that the adjugate of $L(\rho,\lambda(\rho))$, which we write $\mathrm{adj}L$, is made up of cofactors of L and is, therefore, nonzero in $\mathcal{N}_0$. Also

$$(\mathrm{adj}L)L = L(\mathrm{adj}L) = (\det L)I = 0.$$

Hence we can pick out a nonzero row or column from $\mathrm{adj}L$ to obtain a left or right eigenvector, (respectively), of $L(\rho,\lambda(\rho))$ which is also analytic in $\mathcal{N}_0$. (This argument simplifies and extends the discussion of Sun [S1]—see also Theorem 2 of [C1]—but does not obviously extend to multiple eigenvalues.) These observations may be summarised as the following theorem.

THEOREM 2.1. *If $L(\boldsymbol{\rho},\lambda)$ has a simple eigenvalue at $\boldsymbol{\rho}_0$ then there is a neighbourhood $\mathcal{N}_0$ of $\boldsymbol{\rho}_0$ on which there exists an eigenvalue function $\lambda(\boldsymbol{\rho})$ and right and left eigenvector functions that are all analytic functions of $\boldsymbol{\rho}$.*

Note that, if complex eigenvalues and/or eigenvectors are to be considered, then the "complex" version of the implicit function theorem is required in this argument. When the independent variables $\rho_1,\dots,\rho_m$ are real, the corresponding results are obtained by restricting the domain of these independent variables to the real line. In contrast, if $L(\boldsymbol{\rho},\lambda)$ takes values in the real symmetric matrices, then the independent variables, eigenvalues, and eigenvectors are all real valued, and an implicit function theorem for real variables only is appropriate in this case.

Problems with complex eigenparameters (with $\boldsymbol{\rho}$ taking only real values) can also be treated in the context of "real" analysis provided the dependence on λ is polynomial and a suitable linearization with respect to λ is available (see [GLR], for example). The real variable problem may then be formulated by separation of real and imaginary parts.

Now let us consider the case of a multiple unperturbed eigenvalue $\lambda_0 = \lambda(\boldsymbol{\rho}_0)$. The argument leading to Theorem 2.1 breaks down immediately. Indeed, examples of nonanalytic behaviour are easily found, even when $m = 1$. In particular, with $\rho_0 = 0$,

$$\begin{bmatrix} \lambda & \rho \\ \rho^2 & \lambda \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \lambda & 1 \\ \rho & \lambda \end{bmatrix}$$

are of this kind. The unperturbed eigenvalue is semisimple in the first case and not so in the second.

On the other hand, positive examples are also easily constructed. In fact, there is an important class of problems where analytic multiple eigenvalues and analytic eigenvectors can be guaranteed. This is the case of hermitian (or normal) matrix functions of the form

$$L(\rho,\lambda) = A(\rho) - \lambda I$$

with only one *real* parameter, i.e., with $m = 1$. (This is the theory, now classical, of Wolf, Rellich, Kato, and others. See Baumgärtel [B] for a recent comprehensive discussion.)

This beautiful theory for the case $m = 1$ does not extend to $m = 2$ as Rellich's example shows. Take $m = 2, \mathcal{N} \subset \mathbb{R}^2, \boldsymbol{\rho}_0 = (0,0)$, and

$$L(\boldsymbol{\rho},\lambda) = \begin{bmatrix} \rho_1 & \rho_2 \\ \rho_2 & -\rho_1 \end{bmatrix} - \lambda I,$$

and it is easily seen that the eigensurfaces are given by

$$\lambda_+ = (\rho_1^2 + \rho_2^2)^{1/2} \quad \text{and} \quad \lambda_- = -\lambda_+.$$

Neither of these functions is analytic at $\boldsymbol{\rho}_0$.

Nevertheless, we contend that analytic dependence of λ on $\boldsymbol{\rho}$ is realised in sufficiently many cases to make investigation worthwhile under the condition that any multiple eigenvalue to be investigated is semisimple. As noted above, this corresponds to the case in which distinct eigenvalues (or "eigensurfaces") come together at $\boldsymbol{\rho}_0$ and cross one another with no loss of dimension in the linear sum of their corresponding eigenspaces at the common point.

3. Simple and semisimple eigenvalues. It will be useful to describe simple and semisimple eigenvalues in more detail. Since properties of $L(\lambda)$ at a fixed $\boldsymbol{\rho}$ are all that are required here, we suppress $\boldsymbol{\rho}$ in the notations. We use $L^{(r)}(\lambda), r = 1, 2, \ldots,$ for the rth order (partial) derivative of L with respect to λ.

LEMMA 3.1. *Let λ_0 be a semisimple eigenvalue of $L(\lambda)$ and $\mathbf{x} \in \ker L(\lambda_0)$, $\mathbf{x} \neq \mathbf{0}$. Then $L^{(1)}(\lambda_0)\mathbf{x} \notin \operatorname{Im} L(\lambda_0)$.*

Proof. Recall the "local Smith form" for $L(\lambda)$. In a neighbourhood of an eigenvalue λ_0 we have

$$L(\lambda) = E(\lambda)D(\lambda)F(\lambda) \tag{3.1}$$

where

$$D(\lambda) = \operatorname{diag}\left[(\lambda - \lambda_0)^{\alpha_1}, \ldots, (\lambda - \lambda_0)^{\alpha_n}\right], \tag{3.2}$$

$E(\lambda)$ is an analytic matrix function with $\det E(\lambda_0) \neq 0, F(\lambda)$ is a matrix polynomial with $\det F(\lambda)$ equal to a nonzero constant, and $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n (\geq 0)$ are the "partial multiplicities" of λ_0 (see [GR] for more details).

Let $E_0 = E(\lambda_0), F_0 = F(\lambda_0)$, and λ_0 be a semisimple eigenvalue. Thus, for some k, $\alpha_1 = \cdots = \alpha_k = 1$ and $\alpha_{k+1} = \cdots = \alpha_n = 0$. Hence, if D_1 =diag$[I_k, 0]$ and D_2 =diag$[0, I_{n-k}]$, then $L(\lambda_0) = E_0 D_2 F_0$, and $L(\lambda_0)\mathbf{x} = \mathbf{0}$ implies $D_2 F_0 \mathbf{x} = \mathbf{0}$. Thus

$$L^{(1)}(\lambda_0)\mathbf{x} = E_0 D_1 F_0 \mathbf{x} + E_0 D_2 F_0^{(1)} \mathbf{x}.$$

If $L^{(1)}(\lambda_0)\mathbf{x} = L(\lambda_0)\mathbf{u}$ for some $\mathbf{u}$, then

$$D_1 F_0 \mathbf{x} + D_2 F_0^{(1)} \mathbf{x} = D_2 F_0 \mathbf{u},$$

and this implies $D_1 F_0 \mathbf{x} = \mathbf{0}$. Since $D_2 F_0 \mathbf{x} = \mathbf{0}$ also, we conclude that $\mathbf{x} = \mathbf{0}$. Thus, if $L(\lambda_0)\mathbf{x} = \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$, we have $L^{(1)}(\lambda_0)\mathbf{x} \notin \operatorname{Im} L(\lambda_0)$. □

THEOREM 3.2. *The matrix function $L(\lambda)$ has a simple eigenvalue λ_0 if and only if $L(\lambda_0)$ has rank $n - 1$, and $\mathbf{y}^T L^{(1)} \mathbf{x} \neq 0$ where $\mathbf{x}$ and $\mathbf{y}$ are, respectively, right and left eigenvectors of L at λ_0.*

Proof. If λ_0 is a simple eigenvalue, then it is semisimple, and, in (3.1), we have $\alpha_1 = 1, \alpha_2 = \cdots = \alpha_n = 0$, and rank $L(\lambda_0) = n - 1$. If $\mathbf{x}$ is a right eigenvector at λ_0, then, from the lemma, $L^{(1)}(\lambda_0)\mathbf{x} \notin \operatorname{Im} L(\lambda_0)$, and if $\mathbf{y}^T L(\lambda_0) = \mathbf{0}^T, \mathbf{y} \neq \mathbf{0}$, then $\bar{\mathbf{y}}$ is orthogonal to the $(n - 1)$-dimensional subspace $\operatorname{Im} L(\lambda_0)$, but not to the vector $L^{(1)}(\lambda_0)\mathbf{x}$. Thus $\mathbf{y}^T L^{(1)}(\lambda_0)\mathbf{x} \neq 0$.

Conversely, if $L(\lambda_0)$ has rank $n - 1$, then it follows from (3.1) that $\alpha_2 = \cdots = \alpha_n = 0$. If $\alpha_1 \geq 2$, it is found that (with the notation used in the proof of Lemma 3.1)

$$L^{(1)}(\lambda_0) = E_0^{(1)} D_2 F_0 + E_0 D_2 F^{(1)}.$$

Now $L(\lambda_0)\mathbf{x} = \mathbf{0}$ implies $D_2 F_0 \mathbf{x} = \mathbf{0}$, and $\mathbf{y}^T L(\lambda_0) = \mathbf{0}^T$ implies $\mathbf{y}^T E_0 D_2 = \mathbf{0}^T$, and hence, $\mathbf{y}^T L^{(1)}(\lambda_0)\mathbf{x} = 0$. Hence $\mathbf{y}^T L^{(1)}(\lambda_0)\mathbf{x} \neq 0$ implies $\alpha_1 = 1$, and λ_0 is a simple eigenvalue. □

The proof also shows that, for an eigenvalue that is *not* simple, we may have

$$\mathbf{y}^T L^{(1)}(\lambda_0)\mathbf{x} = 0 \tag{3.3}$$

for *all* right eigenvectors $\mathbf{x}$ and left eigenvectors $\mathbf{y}$ (see also Theorem 4.6 of [L]).

4. Normalization of eigenvectors. The result of Theorem 2.1 leaves some freedom for normalization of the eigenvectors while still retaining analytic dependence on $\boldsymbol{\rho}$. Let us first discuss an important but, in some respects, an unfortunate case. It is natural to try to solve equation (1.1),

$$L(\boldsymbol{\rho}, \lambda(\boldsymbol{\rho}))\mathbf{x}(\boldsymbol{\rho}) = \mathbf{0} \quad \text{on } \mathcal{N}, \tag{4.1}$$

simultaneously with the condition on the euclidean norm,

$$\|\mathbf{x}(\boldsymbol{\rho})\|^2 = \mathbf{x}(\boldsymbol{\rho})^*\mathbf{x}(\boldsymbol{\rho}) = 1 \quad \text{on } \mathcal{N}. \tag{4.2}$$

We claim that condition (4.2) is inconsistent with the differentiability of $\mathbf{x}(\boldsymbol{\rho})$ when $\boldsymbol{\rho}$ is complex valued. For example, let $m = 1$ and

$$L(\rho, \lambda) = \begin{bmatrix} \lambda - 1 & -\rho \\ 0 & \lambda - 2 \end{bmatrix}.$$

It is easily seen that $\lambda(\rho) \equiv 2$ is an eigenvalue (for all ρ) with eigenvector $\left[\begin{smallmatrix} \rho \\ 1 \end{smallmatrix}\right]$. This cannot be scaled to satisfy (4.2) and remain analytic. This is because there is no analytic function $f(\rho)$ such that $|f(\rho)|^2 = 1 + |\rho|^2$. A proof of this fact is not difficult to find. This is a problem peculiar to complex analysis and the strong form of differentiation that it employs. If the independent parameters of $\boldsymbol{\rho}$ are real, this problem does not arise (see Case 2 following).

The normalization procedures admitted here retain analyticity of the eigenvector and can all be derived from the following discussion. Let $\hat{\mathbf{x}}(\boldsymbol{\rho})$ be an analytic eigenvector on a neighbourhood $\mathcal{N}_0 \subset \mathbb{C}^m$, and let $\mathbf{z}(\boldsymbol{\rho})^T$ be any row vector analytic on $\mathcal{N}_0$ for which

$$\mathbf{z}(\boldsymbol{\rho}_0)^T\hat{\mathbf{x}}(\boldsymbol{\rho}_0) = 1$$

at a fixed $\boldsymbol{\rho}_0 \in \mathcal{N}_0$. Without loss of generality, we may assume that $\|\hat{\mathbf{x}}(\boldsymbol{\rho}_0)\| = 1$. (Throughout this paper we use the euclidean norm and the induced (spectral) matrix norm.) Now the vector defined by

$$\mathbf{x}(\boldsymbol{\rho}) = \{\mathbf{z}(\boldsymbol{\rho})^T\hat{\mathbf{x}}(\boldsymbol{\rho})\}^{-1}\hat{\mathbf{x}}(\boldsymbol{\rho}) \tag{4.3}$$

is analytic in a neighbourhood $\mathcal{N}$ of $\boldsymbol{\rho}_0, (\mathcal{N} \subset \mathcal{N}_0)$, and also satisfies $\mathbf{x}(\boldsymbol{\rho}_0) = \hat{\mathbf{x}}(\boldsymbol{\rho}_0)$ so that $\|\mathbf{x}(\boldsymbol{\rho}_0)\| = 1$. Furthermore,

$$\mathbf{z}(\boldsymbol{\rho})^T\mathbf{x}(\boldsymbol{\rho}) = 1 \quad \text{on } \mathcal{N}. \tag{4.4}$$

Case 1. Orthogonally constrained eigenvectors. Let $\mathbf{z}$ be a constant nonzero vector function. Then $\mathbf{z}^T\mathbf{x}(\boldsymbol{\rho}) = 1$ on $\mathcal{N}$ implies that, for all partial derivatives $\mathbf{x}^{(r)}(\boldsymbol{\rho}_0)$ of order $r (r \geq 1)$, we have

$$\mathbf{z}^T\mathbf{x}^{(r)}(\boldsymbol{\rho}_0) = 0. \tag{4.5}$$

This is a very convenient orthogonality condition for analytical purposes and is the most commonly used convention in analytic perturbation theory (see Kato [Ka], and §7.1.2 of [B]). Popular choices for $\mathbf{z}$ include the right eigenvector of L at $\boldsymbol{\rho}_0$ (see [TA] for references), or a unit-coordinate vector. The latter may be convenient in numerical work.

The choice of a left eigenvector of L for $\mathbf{z}$ is not recommended. The reason for this is well understood. Even for the classical eigenvalue problem there are situations in which this would result in $\mathbf{z}^T\mathbf{x}(\boldsymbol{\rho}_0)$ being zero or dangerously small. On the other hand, Theorem 3.2 shows that when $\lambda(\boldsymbol{\rho}_0)$ is a simple eigenvalue, $\mathbf{z}^T$ proportional to $\mathbf{y}^T L^{(1)}(\boldsymbol{\rho}_0, \lambda_0)$ is always possible. However, when $\lambda(\boldsymbol{\rho}_0)$ is not semisimple this strategy can also be dangerous (see Theorem 4.6 of [L] and §10 following).

Case 2. Uniformly normed eigenvectors. This case concerns normalization of eigenvectors $\mathbf{x}(\boldsymbol{\rho})$ so that $\|\mathbf{x}(\boldsymbol{\rho})\| = 1$ in a neighbourhood of $\boldsymbol{\rho}_0$. For reasons given above, attention is confined to *real* parameters $\rho_1, \ldots, \rho_m$. In (4.3) we now choose $\mathbf{z}(\boldsymbol{\rho}) = \bar{\mathbf{x}}(\boldsymbol{\rho})$ so that $\mathbf{z}(\boldsymbol{\rho})^T = \mathbf{x}(\boldsymbol{\rho})^*$ and $\|\mathbf{x}(\boldsymbol{\rho})\| \equiv 1$ on $\mathcal{N}$.

Consider the significance of this in terms of Taylor expansions. Without loss of generality, suppose that $\boldsymbol{\rho}_0 = \mathbf{0}$. Let $\mathbf{x}_j, \mathbf{x}_{jk}$ denote derivatives of $\mathbf{x}(\boldsymbol{\rho})$ with respect to ρ_j and ρ_j, ρ_k evaluated at $\boldsymbol{\rho}_0$. (This use of subscripts to denote partial derivatives with respect to ρ_j will also be applied to other dependent variables.) Consider the expansion of $\mathbf{x}(\boldsymbol{\rho})$ about $\boldsymbol{\rho} = \mathbf{0}$ in terms of ρ_j and ρ_k. We have

$$\mathbf{x}(\rho_j, \rho_k) = \mathbf{x}_0 + \rho_j \mathbf{x}_j + \rho_k \mathbf{x}_k + \frac{1}{2}(\rho_j^2 \mathbf{x}_{jj} + 2\rho_j\rho_k \mathbf{x}_{jk} + \rho_k^2 \mathbf{x}_{kk}) + O(\|\boldsymbol{\rho}\|^3). \tag{4.6}$$

The normalization $\|\mathbf{x}(\boldsymbol{\rho})\| \equiv 1$ on $\mathcal{N}$ means that $\|\mathbf{x}_0\| = 1$ and, on examining coefficients in $\mathbf{x}(\boldsymbol{\rho})^*\mathbf{x}(\boldsymbol{\rho}) = 1$, it is deduced from (4.6) that, for $1 \le j, k \le m$,

$$\mathbf{x}_0^*\mathbf{x}_j + \mathbf{x}_j^*\mathbf{x}_0 = 0 \tag{4.7}$$

and

$$\mathbf{x}_0^*\mathbf{x}_{jk} + \mathbf{x}_{jk}^*\mathbf{x}_0 = -(\mathbf{x}_j^*\mathbf{x}_k + \mathbf{x}_k^*\mathbf{x}_j). \tag{4.8}$$

These are necessary conditions for normalization of this kind.

When the eigenvector is real these necessary conditions reduce to

$$\mathbf{x}_0^T\mathbf{x}_j = 0, \qquad \mathbf{x}_0^T\mathbf{x}_{jk} = -\mathbf{x}_j^T\mathbf{x}_k. \tag{4.9}$$

We make one more remark about the euclidean normalization. It should not be imagined that, when $\mathbf{x}(\boldsymbol{\rho})$ is analytic and $\|\mathbf{x}(\boldsymbol{\rho})\| \equiv 1$, the magnitude of the derivative of $\mathbf{x}(\boldsymbol{\rho})$ is also controlled. For example, if $m = 1, \alpha \in \mathbb{R}$, and $\mathbf{x}(\rho) \in \mathbb{R}^2$ is defined by

$$x_1(\rho) = \left(\frac{1 - \rho - \alpha^2\rho^2}{1 - \rho}\right)^{1/2}, \qquad x_2(\rho) = \frac{\alpha\rho}{(1-\rho)^{1/2}},$$

then $\|\mathbf{x}(\rho)\| \equiv 1$ and, as $|\rho| \to 0$,

$$\mathbf{x}(\rho) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \rho \begin{bmatrix} 0 \\ \alpha \end{bmatrix} + O(|\rho|^2).$$

Thus, the norm of the derivatives of $\mathbf{x}$ at $\rho = 0$ is $|\alpha|$ and may be arbitrarily large, even though $\|\mathbf{x}(\rho)\| \equiv 1$.

Case 3. Left normalized eigenvectors. Let $\mathbf{y}(\boldsymbol{\rho})$ be an analytic eigenvector for the transposed matrix function. Thus, on a neighbourhood of $\boldsymbol{\rho}_0$,

$$L(\boldsymbol{\rho}, \lambda(\boldsymbol{\rho}))^T\mathbf{y}(\boldsymbol{\rho}) = \mathbf{0}. \tag{4.10}$$

From Theorem 3.2, when $\lambda(\rho_0)$ is a simple eigenvalue of L, then at $(\rho_0, \lambda(\rho_0))$, we may assume that

$$\mathbf{y}^T L^{(1)} \hat{\mathbf{x}} = 1 \tag{4.11}$$

for some right eigenvector $\hat{\mathbf{x}}$ of norm one. In this case the choice

$$\mathbf{z}(\rho)^T = \mathbf{y}(\rho)^T L^{(1)}(\rho, \lambda(\rho)) \tag{4.12}$$

in (4.3) may offer advantages and leads to another special case of (4.4),

$$\mathbf{y}(\rho)^T L^{(1)}(\rho, \lambda(\rho))\mathbf{x}(\rho) \equiv 1 \tag{4.13}$$

on $\mathcal{N}$.

In the case of the classical eigenvalue problem this condition can be written $\mathbf{y}(\rho)^T\mathbf{x}(\rho) \equiv 1$.

Our investigations will focus primarily on Cases 1 and 2.

5. Bordered-matrix equations and their solutions. The discussion of §§2 and 4 leads us to the study of equations of the form

$$\begin{cases} L(\rho, \lambda(\rho))\mathbf{x}(\rho) = \mathbf{0}, \\ \mathbf{z}(\rho)^T\mathbf{x}(\rho) = 1, \end{cases} \tag{5.1}$$

when $\lambda, \mathbf{x}$, and $\mathbf{z}$ depend analytically on ρ in a neighbourhood $\mathcal{N}$ of some ρ_0. Differentiating both equations with respect to ρ_j yields

$$L\mathbf{x}_j + \lambda_j L^{(1)}\mathbf{x} = -L_j\mathbf{x} \tag{5.2}$$

and $\mathbf{z}^T\mathbf{x}_j = -\mathbf{z}_j^T\mathbf{x}$. These two equations can be combined in the form

$$\begin{bmatrix} L & L^{(1)}\mathbf{x} \\ \mathbf{z}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_j \\ \lambda_j \end{bmatrix} = \begin{bmatrix} -L_j\mathbf{x} \\ -\mathbf{z}_j^T\mathbf{x} \end{bmatrix}. \tag{5.3}$$

In the case of orthogonally constrained eigenvectors this is an equation for $\mathbf{x}_j, \lambda_j$ in terms of $\mathbf{x}, \mathbf{z}$, and $\mathbf{z}_j$. The coefficient matrix is obtained by a suitable "bordering" of the singular matrix L. For this analysis it is important to note that the assumption of analytic solutions to (5.1) ensures the *consistency* of the linear system (5.3) for $\mathbf{x}_j$ and λ_j.

When considering algorithms for eigenvalues and eigenvectors themselves (rather than their derivatives) it should be noted that the coefficient matrix of (5.3) arises in the implementation of Newton's method; with the exception that it is evaluated at the current approximation for an eigenvalue, rather than at the eigenvalue itself (see [KK] or [Ru], for example). Linear systems with a bordered coefficient matrix, as in (5.3), play a prominent role in the study of continuation methods for fully nonlinear eigenvalue problems (see [Ch], [Kel], and [Rh], for example). It is interesting to make this connection precise, but in the interests of continuity this is postponed to the Appendix.

The first step in the analysis of equations such as (5.3) (with any term on the right) is the following lemma.

LEMMA 5.1. *Let $\lambda_0 = \lambda(\rho_0)$ be a semisimple eigenvalue of $L(\rho, \lambda)$. Then if* $\mathbf{x} \in \ker L(\lambda_0)$ *with* $\mathbf{x} \neq \mathbf{0}$,

$$\ker \begin{bmatrix} L(\lambda_0) & L^{(1)}(\lambda_0)\mathbf{x} \end{bmatrix} = \left\{ \begin{bmatrix} \mathbf{s} \\ 0 \end{bmatrix} : \mathbf{s} \in \ker L(\lambda_0) \right\}.$$

Proof. If $[\mathbf{s}^T \ \alpha]^T$ is in the kernel of $[L \quad L^{(1)}\mathbf{x}]$, then

$$\alpha L^{(1)}\mathbf{x} = -L\mathbf{s}.$$

However, if λ_0 is semisimple this contradicts Lemma 3.1 unless $\alpha = 0$, in which case $\mathbf{s} \in \ker L$. □

The following theorem includes results of Tan, Andrew, and de Hoog (see [T] and [AHT]) concerning classical eigenvalue problems. For convenience, let $\mathcal{Z}$ denote the subspace of all vectors orthogonal to $\bar{\mathbf{z}}$, i.e., vectors $\mathbf{x}$ for which $\bar{\mathbf{z}}^*\mathbf{x} = \mathbf{z}^T\mathbf{x} = 0$.

THEOREM 5.2. *Let $\lambda_0 = \lambda(\rho_0)$ be a semisimple eigenvalue of $L(\rho, \lambda)$, $\mathbf{x} \in \ker L(\lambda_0)$ with $\mathbf{x} \neq \mathbf{0}$, and assume that the equation*

$$\begin{bmatrix} L & L^{(1)}\mathbf{x} \\ \mathbf{z}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{t} \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbf{g} \\ \gamma \end{bmatrix} \tag{5.4}$$

for given $\mathbf{g} \in \mathbb{C}^n, \gamma \in \mathbb{C}$ is consistent. If $\mathbf{t}_p, \beta_p$ define a particular solution, then the general solution has the form

$$\mathbf{t} = \mathbf{t}_p + \mathbf{s}, \qquad \beta = \beta_p,$$

where $\mathbf{s} \in (\ker L) \cap \mathcal{Z}$.

Furthermore, when $\mathbf{z}^T\mathbf{x} = 1$ and $\ker L$ *has dimension r, the dimension of* $(\ker L) \cap \mathcal{Z}$ *is $r - 1$.*

Proof. Let $\mathbf{t}_p, \beta_p$ define a particular solution of (5.4). Then the general solution of (5.4) is defined by pairs $\mathbf{t}_p + \mathbf{s}, \beta_p + \beta$, where

$$\begin{bmatrix} L & L^{(1)}\mathbf{x} \\ \mathbf{z}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{s} \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}.$$

Using Lemma 5.1 this means that $\beta = 0$ and $\mathbf{s} \in (\ker L) \cap \mathcal{Z}$.

The relation $\mathbf{z}^T\mathbf{x} = \mathbf{x}^*\bar{\mathbf{z}} = 1$ implies that $\mathbf{x} \notin \mathcal{Z}$. Since $\mathbf{x} \in \ker L$ we have the *proper* inclusion

$$(\ker L) \cap \mathcal{Z} \subset \ker L.$$

Since $\mathcal{Z}$ has dimension $n-1$ it follows that the intersection has dimension $r-1$. □

COROLLARY 5.3. *Let $L(\rho, \lambda)$ have a semisimple eigenvalue λ_0 at ρ_0 and let $\lambda(\rho)$, $\mathbf{x}(\rho)$ be analytic functions for which* (5.1) *is satisfied in a neighbourhood $\mathcal{N}$ of ρ_0. Then the set of all solutions of*

$$\begin{bmatrix} L & L^{(1)}\mathbf{x} \\ \mathbf{z}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{t} \\ \beta \end{bmatrix} = - \begin{bmatrix} L_j\mathbf{x} \\ \mathbf{z}_j^T\mathbf{x} \end{bmatrix} \tag{5.5}$$

is given by

$$\mathbf{t} = \mathbf{x}_j + \mathbf{s}, \qquad \beta = \lambda_j, \tag{5.6}$$

where $\mathbf{s} \in (\ker L) \cap \mathcal{Z}$.

Proof. Since $\lambda(\boldsymbol{\rho}), \mathbf{x}(\boldsymbol{\rho})$ are analytic, they are differentiable and equation (5.3) holds. This shows that equation (5.5) is consistent with the particular solution $\mathbf{t} = \mathbf{x}_j, \beta = \lambda_j$. Now apply the theorem. □

Equation (5.3) also applies under weaker hypotheses on $\lambda(\boldsymbol{\rho})$ and $\mathbf{x}(\boldsymbol{\rho})$. If we admit variation of only one scalar parameter at a time and suppose that $\lambda(\boldsymbol{\rho}_0)$ is a multiple eigenvalue, then $\lambda(\rho_j)$ has an expansion in powers of $(\rho_j - \rho_{j,0})^{1/k}$ about $\rho_{j,0}$ for some positive integer k. But without further hypotheses, k continuous eigenvector functions need not exist at $\rho_{j,0}$. With the hypothesis of a *semisimple* eigenvalue at $\rho_{j,0}$ the situation improves to the extent that: (a) the expansion for $\lambda(\rho_j)$ in fractional powers takes the form

$$\lambda(\rho_j) = \lambda_0 + (\rho_j - \rho_{j,0})\lambda_j + O(|\rho_j - \rho_{j,0}|^{1+1/k})$$

as $\rho_j \to \rho_{j,0}$, and (b) there are k eigenfunctions with fractional power expansions

$$\mathbf{x}_0 + (\rho_j - \rho_{j,0})^{1/k}\mathbf{x}^{(1)} + (\rho_j - \rho_{j,0})^{2/k}\mathbf{x}^{(2)} + \cdots$$

in a neighbourhood of $\rho_{j,0}$. (See Theorem 11.7.1 of [LT] and Theorem II.2.3 of [Ka], for example.) When $\boldsymbol{\rho}$ is real, it follows that, in this case, first partial derivatives of $\lambda(\boldsymbol{\rho})$ exist at $\boldsymbol{\rho}_0$, although this is not generally the case for $\mathbf{x}(\boldsymbol{\rho})$. When $\boldsymbol{\rho}$ is complex, λ_j in the above expansion for $\lambda(\rho_j)$ may still be interpreted as a "partial derivative" of $\lambda(\boldsymbol{\rho})$ at $\boldsymbol{\rho}_0$, and the term "partial derivatives" in Corollary 5.4 is to be interpreted in this weak sense. However, with appropriate hypotheses, a different version of the last corollary is obtained which may extend its applicability even though the hypotheses are still not easily verified. Note first of all that, when all necessary partial derivatives exist, (5.3) is satisfied. Then we obtain the following corollary.

COROLLARY 5.4. *Let $L(\boldsymbol{\rho}, \lambda)$ have a semisimple eigenvalue λ_0 at $\boldsymbol{\rho}_0$. Then $\lambda(\boldsymbol{\rho})$ has partial derivatives λ_j at $\boldsymbol{\rho}_0, j = 1, 2, \ldots, m$. If, in addition, $\mathbf{x}(\boldsymbol{\rho})$ and $\mathbf{z}(\boldsymbol{\rho})$ have first-order partial derivatives at $\boldsymbol{\rho}_0$, then the conclusions of Corollary* 5.3 *hold.*

COROLLARY 5.5. *If, in Corollary* 5.3, *λ_0 is a simple eigenvalue and $\mathbf{z}^T\mathbf{x} = 1$, then the coefficient matrix in* (5.5) *is nonsingular and the unique solution is $\mathbf{t} = \mathbf{x}_j, \beta = \lambda_j$.*

Proof. In Theorem 5.2 we have $r = 1$ so that $(\ker L) \cap \mathcal{Z} = \{\mathbf{0}\}$ and (5.5) has the unique solution $\mathbf{t} = \mathbf{x}_j, \beta = \lambda_j$. □

Note that, in Corollary 5.5, it is not necessary to *assume* the existence of analytic functions $\lambda(\boldsymbol{\rho})$ and $\mathbf{x}(\boldsymbol{\rho})$ as in Corollary 5.3. This is assured by Theorem 2.1.

Corollary 5.5 may be seen as a central result of this development. From the point of view of numerical computation the possibility of finding derivatives λ_j and $\mathbf{x}_j$ simply by solving a nonsingular linear system is attractive. Of course, it is assumed that the eigenvalue λ_0 itself and an eigenvector $\mathbf{x}_0$ (and possibly a left eigenvector $\mathbf{y}_0$) are already known to a sufficient accuracy. In this context, the role of Corollaries 5.3 and 5.4 is to provide some insight into the behaviour of algorithms designed for the nonsingular case in the neighbourhood of a multiple semisimple eigenvalue.

Now consider the normalization question in the context of (5.5). In the case of orthogonally constrained eigenvectors, $\mathbf{z}$ is independent of $\boldsymbol{\rho}$ so that $\mathbf{z}_j = \mathbf{0}$ in (5.5).

The analysis of uniformly normed eigenvectors is more delicate when the eigenvector is complex. In this case $\mathbf{z}(\boldsymbol{\rho})^T = \mathbf{x}(\boldsymbol{\rho})^*$ (recall that $\boldsymbol{\rho} \in \mathbb{R}^m$ in this case so that $\mathbf{z}_j^T = \mathbf{x}_j^*$ in (5.5)), and this "unknown" appears on both the left and right of (5.5) (see also (4.7)). However, if we find an $\mathbf{x}_j$ for which $\mathbf{x}_0^*\mathbf{x}_j = 0$, as in Case 1, then the vector $\mathbf{x}_j$ certainly satisfies (4.7) and (5.5). But now there is a family of solutions of (5.5) for which $\mathbf{t}$ belongs to the set

$$\{\mathbf{x}_j + i\alpha\mathbf{x}_0 \colon \alpha \in \mathbb{R}\}. \tag{5.7}$$

Uniqueness can be imposed by applying a further condition on $\mathbf{x}_j$. For example, if it is required that $\mathbf{x}_j$ be orthogonal to $\mathbf{x}_0$, we obtain the unique $\mathbf{x}_j$ of Case 1. A second possibility is minimization of the norms of the derivative vectors of (5.7). Thus, determine an "optimal" derivative vector $\hat{\mathbf{x}}_j$ by

$$\|\hat{\mathbf{x}}_j\| = \min_{\alpha} \|\mathbf{x}_j + i\alpha\mathbf{x}_0\|. \tag{5.8}$$

But, as $\mathbf{x}_j$ and $\mathbf{x}_0$ are othogonal, it is found that $\alpha = 0$.

When the eigenvector is real, (4.7) reduces to the first equation of (4.9) and the derivative vector is the same as in Case 1. Thus we have the following proposition.

PROPOSITION 5.6. *Let λ_0 be a simple eigenvalue. If, in* (5.5), $\mathbf{z}^T = \mathbf{x}_0^*$ *and* $\mathbf{z}_j = \mathbf{0}$, *then for the unique solution* $\mathbf{t} = \hat{\mathbf{x}}_j$, $\beta = \lambda_j$, *we have*:

(a) $\hat{\mathbf{x}}_j$ *is the derivative of an orthogonally constrained eigenvector function*; *and*

(b) $\hat{\mathbf{x}}_j$ *is the derivative of a uniformly normed eigenvector function which, when the eigenvector takes complex values, also satisfies* (5.8).

6. Higher-order derivatives. It has been observed elsewhere (see [CMa], for example) that, if $\mathbf{x}_j$ and λ_j are determined by equations of the form (5.3), then higher-order derivatives satisfy equations with the same coefficient matrix. All the terms in that equation can be considered as analytic functions of $\boldsymbol{\rho}$ and so may be differentiated with respect to $\rho_k, 1 \le k \le m$. A little calculation shows that

$$\begin{bmatrix} L & L^{(1)}\mathbf{x}_0 \\ \mathbf{z}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_{jk} \\ \lambda_{jk} \end{bmatrix} = \begin{bmatrix} -\mathbf{b}_1 \\ -b_2 \end{bmatrix}, \tag{6.1}$$

where
(6.2)

$$\mathbf{b}_1 = (L_k + \lambda_k L^{(1)})\mathbf{x}_j + (L_j + \lambda_j L^{(1)})\mathbf{x}_k + (L_{jk} + \lambda_j L_k^{(1)} + \lambda_k L_j^{(1)} + \lambda_j\lambda_k L^{(2)})\mathbf{x}_0,$$
$$b_2 = \mathbf{z}_k^T\mathbf{x}_j + \mathbf{z}_j^T\mathbf{x}_k + \mathbf{z}_{jk}^T\mathbf{x}_0.$$

In particular, when $k = j$,

$$\begin{aligned} \mathbf{b}_1 &= 2(L_j + \lambda_j L^{(1)})\mathbf{x}_j + (L_{jj} + 2\lambda_j L_j^{(1)} + \lambda_j^2 L^{(2)})\mathbf{x}_0, \\ b_2 &= 2\mathbf{z}_j^T\mathbf{x}_j + \mathbf{z}_{jj}^T\mathbf{x}_0. \end{aligned} \tag{6.3}$$

As in §5, the analytic property of $\lambda(\boldsymbol{\rho}), \mathbf{x}(\boldsymbol{\rho})$ ensures both the existence of the derivatives and the consistency of equations (6.1). Furthermore, Theorem 5.2 then gives a complete description of the solution set of (6.1).

As in the case of Corollary 5.4, it is easily seen that (6.1) holds under weaker assumptions. For example, suppose that expansions such as (4.6) are known to hold (for both $\mathbf{x}$ and λ), but with $\| \boldsymbol{\rho} \|^3$ replaced by $\| \boldsymbol{\rho} \|^{2+1/p}$ for some integer $p > 0$. Equation (6.1) can then be obtained by first expanding L in (1.1) using hypothesis (i), then substituting the appropriate expansions for λ and $\mathbf{x}$ and then comparing coefficients.

Now consider the options for normalizing the eigenvector discussed in §4.

Orthogonally constrained eigenvector. In this case $\mathbf{z}$ is a constant vector and so $b_2 = 0$ in (6.1).

Complex uniformly normed eigenvector. Here we have $\mathbf{z}(\boldsymbol{\rho})^T = \mathbf{x}(\boldsymbol{\rho})^*$ in (6.1) and this equation includes the necessary condition (4.8). As with the first derivative vector, $\mathbf{x}_{jk}$ is not uniquely determined by these conditions. If $\mathbf{x}_{jk}$ is one solution, then the set of all solutions is found to have the form

$$\{\mathbf{x}_{jk} + i\alpha\mathbf{x}_0 \colon \alpha \in \mathbb{R}\}. \tag{6.4}$$

Since the right-hand side of (4.8) is real, one solution of that equation is obtained by finding $\mathbf{x}_{jk}$ such that

$$\mathbf{x}_0^*\mathbf{x}_{jk} = -\frac{1}{2}(\mathbf{x}_j^*\mathbf{x}_k + \mathbf{x}_k^*\mathbf{x}_j).$$

With this choice of $\mathbf{x}_{jk}$ the vector in the set of (6.4) of minimal norm is easily seen to be given by taking $\alpha = 0$.

Real uniformly normed eigenvector. The problem of nonuniqueness no longer arises. In view of the second equation of (4.9), equation (6.1) is to be solved with $\mathbf{z}^T = \mathbf{x}_0^*$ and $b_2 = \mathbf{x}_j^T\mathbf{x}_k$. These results may be summarised as the following proposition.

PROPOSITION 6.1. *Let λ_0 be a simple eigenvalue.*

(a) *If, in* (6.1), $\mathbf{z}^T = \mathbf{x}_0^*$, $\mathbf{b}_1$ *is given by* (6.2) *and* $b_2 = 0$, *then the unique solution* $\mathbf{x}_{jk}$ *is a second-order derivative of an orthogonally constrained eigenvector function.*

(b) *If, in* (6.1), $\mathbf{z}^T = \mathbf{x}_0^*$, $\mathbf{b}_1$ *is given by* (6.2) *and* $b_2 = \frac{1}{2}(\mathbf{x}_j^*\mathbf{x}_k + \mathbf{x}_k^*\mathbf{x}_j)$, *then the unique solution* $\hat{\mathbf{x}}_{jk}$ *is a second-order derivative of a uniformly normed eigenvector function which, when the eigenvector takes complex values, has minimum norm in the class* (6.4).

For the important case of uniformly normed complex eigenvectors, Propositions 5.6(b) and 6.1(b) initiate a general strategy for choosing a derivative vector, say $\mathbf{x}_{k_1\ldots k_m}$ (differentiated k_j times with respect to ρ_j); namely, we choose that candidate of minimum euclidean norm. It remains to be shown that this choice does, indeed, generate an *analytic* vector function $\mathbf{x}(\boldsymbol{\rho})$. The hypothesis is that there is an analytic function

$$\mathbf{x}(\boldsymbol{\rho}) = \sum_{k_1,\ldots,k_m \geq 0} \mathbf{x}_{k_1\ldots k_m}(\rho_1 - \rho_1^0)^{k_1} \ldots (\rho_m - \rho_m^0)^{k_m}$$

defined on a neighbourhood of $\boldsymbol{\rho}_0 = (\rho_1^0, \ldots, \rho_m^0)$. In other words (see [BM]), the n scalar series

$$x^j(\boldsymbol{\rho}) = \sum x^j_{k_1\ldots k_m}(\rho_1 - \rho_1^0)^{k_1} \ldots (\rho_m - \rho_m^0)^{k_m}, \qquad j = 1, 2, \ldots, n,$$

are absolutely convergent in a polycylinder of $\mathbb{C}^m$. This means that

$$|x^j_{k_1\ldots k_m}| \leq \frac{M}{R_1^{k_1} \ldots R_m^{k_m}}$$

for some $R_1 > 0, \ldots, R_m > 0$, and $j = 1, 2, \ldots, n$. These inequalities imply

$$\|\mathbf{x}_{k_1\ldots k_m}\| \leq \frac{nM}{R_1^{k_1} \ldots R_m^{k_m}}.$$

Now our strategy determines a derivative vector $\hat{\mathbf{x}}_{k_1 \dots k_m}$ with

$$\|\hat{\mathbf{x}}_{k_1 \dots k_m}\| \le \|\mathbf{x}_{k_1 \dots k_m}\| \le \frac{nM}{R_1^{k_1} \dots R_m^{k_m}}$$

whence

$$|\hat{x}^j_{k_1 \dots k_m}| \le \frac{nM}{R_1^{k_1} \dots R_m^{k_m}}.$$

As n is fixed, it follows that

$$\hat{\mathbf{x}}(\boldsymbol{\rho}) = \sum \hat{\mathbf{x}}_{k_1 \dots k_m} (\rho_1 - \rho_1^0)^{k_1} \dots (\rho_m - \rho_m^0)^{k_m}$$

is also absolutely convergent on the polycylinder defined by $R_1, \dots, R_m$.

Thus, consistent use of the minimization strategy initiated in Propositions 5.6(b) and 6.1(b) will determine a uniformly normed analytic eigenvector function $\mathbf{x}(\boldsymbol{\rho})$ on a neighbourhood of $\boldsymbol{\rho}_0$.

7. Formulae for derivatives using the Moore–Penrose inverse. We operate always under conditions for which (5.3) holds. From this equation an explicit formula for the derivative λ_j is obtained by introducing a left eigenvector $\mathbf{y}_0$ for L for which

$$\mathbf{y}_0^T L^{(1)} \mathbf{x}_0 \neq 0. \tag{7.1}$$

Thus

$$\lambda_j = -\frac{\mathbf{y}_0^T L_j \mathbf{x}_0}{\mathbf{y}_0^T L^{(1)} \mathbf{x}_0}. \tag{7.2}$$

As noted in Theorem 3.2, condition (7.1) is automatically satisfied when λ is a simple eigenvalue. Furthermore, when λ is multiple and semisimple, such a left eigenvector can always be found. From the point of view of analysis it is therefore convenient to treat λ_j as known in (5.2). With this understanding the solutions of (5.2),

$$L\mathbf{x}_j = -(\lambda_j L^{(1)} + L_j)\mathbf{x}_0 \tag{7.3}$$

can be expressed in terms of generalized inverses of L. In particular, let L^+ denote the Moore–Penrose generalized inverse of L (see [BG], [CMe], or [LT], for example), and recall that L^+L is the orthogonal projector on ImL^* along ker$L$ and LL^+ is the orthogonal projector on ImL along kerL^*.

The general solution of (7.3), when λ is a simple eigenvalue, has the form

$$\mathbf{x}_j = -L^+(\lambda_j L^{(1)} + L_j)\mathbf{x}_0 + \alpha \mathbf{x}_0, \tag{7.4}$$

where α is an arbitrary scalar. The choice $\alpha = 0$ determines a vector $\mathbf{x}_j$ in Im$L^* = (\text{ker}L)^\perp$. Thus, the derivative vector is orthogonal to $\mathbf{x}$. This is the solution determined by putting $\mathbf{z}^T = \mathbf{x}_0^*$ in the system (5.3). Other choices of α correspond to other choices of the normalization vector $\mathbf{z}$. Clearly, the choice $\alpha = 0$ produces the vector $\mathbf{x}_j$ of minimum norm in the manifold of solutions of (7.3).

An explicit formula for general $\mathbf{z}$ is contained in the next theorem.

THEOREM 7.1. *Let L have a simple eigenvalue λ at ρ_0 and corresponding eigenvector $\mathbf{x}_0$ with $\|\mathbf{x}_0\| = 1$. Let $\mathbf{z}$ in (5.3) be a constant vector for which $\mathbf{z}^T\mathbf{x}_0 = 1$. Then the unique solution of (5.3) is given by (7.2) and*

$$\mathbf{x}_j = -(I - \mathbf{x}_0\mathbf{z}^T)L^+(\lambda_j L^{(1)} + L_j)\mathbf{x}_0. \tag{7.5}$$

When $\mathbf{z}^T = \mathbf{x}_0^*$, we have

$$\mathbf{x}_j = -L^+(\lambda_j L^{(1)} + L_j)\mathbf{x}_0, \tag{7.6}$$

and this is where $\min_{\mathbf{z}} \|\mathbf{x}_j\|$ is attained.

Proof. Note that $I - \mathbf{x}_0\mathbf{z}^T$ is the (generally skew) projector on $\mathcal{Z}$ along ker L. Here, as in §5, $\mathcal{Z} = \{\mathbf{y}\colon \mathbf{z}^T\mathbf{y} = 0\}$, the orthogonal complement of span $\{\bar{\mathbf{z}}\}$. Equation (5.3) implies that $\mathbf{x}_j \in \mathcal{Z}$, so multiplying (7.4) on the left by $I - \mathbf{x}_0\mathbf{z}^T$ we immediately obtain (7.5). Clearly, we have $\mathbf{z}^T\mathbf{x}_j = 0$ as required by (5.3). Because $\mathrm{Im}L^+ = \mathrm{Im}L^*$ and $\mathbf{x}_0 \in \ker L = (\mathrm{Im}L^*)^\perp = (\mathrm{Im}L^+)^\perp$, we have $\mathbf{x}_0^* L^+ = \mathbf{0}^T$. Thus, choosing $\mathbf{z}^T = \mathbf{x}_0^*$ in (7.5) results in (7.6). □

In the context of normalization strategies discussed in §4, this result provides an argument for using $\mathbf{z}^T = \mathbf{x}_0^*$ in (5.3). It is the normalization (of Case 1) producing the derivative vector of least norm, i.e., this determines an eigenvector of least sensitivity.

Let $\mu = \|L^+(\lambda_j L^{(1)} + L_j)\mathbf{x}_0\|$, the least norm attainable for $\mathbf{x}_j$. Then for other choices of $\mathbf{z}$ (7.5) gives

$$\|\mathbf{x}_j\| \le \mu\|I - \mathbf{x}_0\mathbf{z}^T\|.$$

Note that $\|\mathbf{z}\| \ge 1$. Also, a standard argument, using the easily proved fact that any nonzero eigenvalues of $(I - \mathbf{x}_0\mathbf{z}^T)^*(I - \mathbf{x}_0\mathbf{z}^T) - I - \bar{\mathbf{z}}\,\mathbf{z}^T$ must be $-1 \pm \|\mathbf{z}\|$, shows that

$$\|\mathbf{z}\|^{\frac{1}{2}} \le \|I - \mathbf{x}_0\mathbf{z}^T\| \le \|\mathbf{z}\|(1 + \|\mathbf{z}\|^{-1})^{\frac{1}{2}}. \tag{7.7}$$

These inequalities indicate that, when the cosine of the angle between $\bar{\mathbf{z}}$ and $\mathbf{x}_0$ is *small*, so that $\|\mathbf{z}\|$ must be *large* to maintain the condition $\mathbf{z}^T\mathbf{x}_0 = 1$, then $\|\mathbf{x}_j\|$ may be large of the same order.

A lower bound for $\|\mathbf{x}_j\|$ can be obtained by using the singular value decomposition (SVD) of the constant matrix $L(= L(\rho_0, \lambda(\rho_0)))$. Let

$$D = \mathrm{diag}\ [\sigma_1, \dots, \sigma_{n-1}, 0] \tag{7.8}$$

be the diagonal matrix of singular values of L and U, V be the unitary matrices of singular vectors (see §5.7 of [LT], for example). Thus, $L = UDV^*$ and $L^+ = VD^+U^*$, where

$$D^+ = \mathrm{diag}[\sigma_1^{-1}, \dots, \sigma_{n-1}^{-1}, 0].$$

As in Theorem 7.1 we assume that $L(\lambda)$ has a simple zero at ρ_0, and this implies that $\sigma_1, \dots, \sigma_{n-1}$ are nonzero. Note also that $\mathbf{x}_0$ has the same direction as $\mathbf{v}_n$, where $V = [\mathbf{v}_1, \dots, \mathbf{v}_n]$.

For any $k \in \{1, 2, \dots, n-1\}$ we have $\mathbf{v}_k^*\mathbf{x}_0 = 0$ and

$$\mathbf{v}_k^* L^+ = \sigma_k^{-1}\mathbf{u}_k^*.$$

Multiplying (7.5) on the left by $\mathbf{v}_k^*$ we obtain

$$\mathbf{v}_k^*\mathbf{x}_j = -\sigma_k^{-1}\mathbf{u}_k^*(\lambda_j L^{(1)} + L_j)\mathbf{x}_0,$$

and by the Schwarz inequality

$$\frac{|\mathbf{u}_k^*(\lambda_j L^{(1)} + L_j)\mathbf{x}_0|}{\sigma_k} \le \|\mathbf{x}_j\|. \tag{7.9}$$

THEOREM 7.2. *Given the hypotheses of Theorem* 7.1

(7.10)
$$\max_{1\le k\le n-1} \frac{|\mathbf{u}_k^*(\lambda_j L^{(1)} + L_j)\mathbf{x}_0|}{\sigma_k} \le \|\mathbf{x}_j\| \le \|\mathbf{z}\|(1+\|\mathbf{z}\|^{-1})^{\frac{1}{2}}\|L^+(\lambda_j L^{(1)} + L_j)\mathbf{x}_0\|.$$

Proof. The lower bound follows from (7.9). The upper bound follows from (7.5) and (7.7). □

Note that the terms in $\|\mathbf{z}\|$ on the right can be replaced by "one" when $\mathbf{z}^T = \mathbf{x}_0^*$.

Consider the significance of Theorem 7.2 for computation in the neighbourhood of a multiple eigenvalue. Thus suppose that at $\boldsymbol{\rho}_1$ there is an eigenvalue $\lambda_1 = \lambda(\boldsymbol{\rho}_1)$ of multiplicity $\alpha \ge 2$. Guided by the discussion of §2 we suppose that λ_1 is a semisimple eigenvalue of $L(\boldsymbol{\rho}, \lambda)$, that $\lambda(\boldsymbol{\rho})$ is analytic in a neighbourhood $\mathcal{N}$ of $\boldsymbol{\rho}_1$, and, when $\boldsymbol{\rho} \in \mathcal{N}$ and $\boldsymbol{\rho} \ne \boldsymbol{\rho}_1$, then $\boldsymbol{\lambda}(\boldsymbol{\rho})$ is a simple eigenvalue.

The inequalities (7.10) apply to $\lambda(\boldsymbol{\rho})$ when $\boldsymbol{\rho} \ne \boldsymbol{\rho}_1$ and they suggest that, as $\boldsymbol{\rho} \to \boldsymbol{\rho}_1, \alpha - 1$ singular values σ_k will converge to zero, so that $\|\mathbf{x}_j\|$ will be unbounded on $\mathcal{N}$. That this is not necessarily the case is shown by observing that the consistency of (5.2) gives $(\lambda_j L^{(1)} + L_j)\mathbf{x} \in \mathrm{Im} L$ for all $\boldsymbol{\rho} \in \mathcal{N}$ and $\mathbf{x} \in \ker L$.

Thus, since $\mathbb{C}^n = \mathrm{Im} L \oplus \ker L^*$, we have

(7.11)
$$\mathbf{u}^*(\lambda_j L^{(1)} + L_j)\mathbf{x} = 0,$$

whenever $\mathbf{u} \in \ker L^*$. Furthermore, if $\sigma_k(\boldsymbol{\rho}_1) = 0$ in (7.10), then we *also* have $\mathbf{u}_k \in \ker L^*$ and (7.11) will apply.

Similarly, for any $\boldsymbol{\rho} \in \mathcal{N}$ we have SVDs $L = UDV^*$ and $L^+ = VD^+U^*$, where

$$D = \mathrm{diag}\ [\sigma_1, \sigma_2, \dots, 0], \qquad D^+ = \mathrm{diag}[\sigma_1^{-1}, \sigma_2^{-1}, \dots, 0],$$

and $\|L^+\|$ will be unbounded on $\mathcal{N}$. However, the orthogonality (7.11) applies once more to show that the vector $L^+(\lambda_j L^{(1)} + L_j)\mathbf{x}$ (appearing on the right on (7.10)) may be bounded on $\mathcal{N}$.

It should also be recognized, however, that computational errors in finding $\mathbf{x}, \lambda_j$ and $(\lambda_j L^{(1)} + L)\mathbf{x}$ will generally satisfy no orthogonality condition and so lead to some numerical instability.

8. Formulae for derivatives using the group inverse. There are some situations in which the use of the group inverse is natural in the formulation of explicit formulae for derivatives of the eigenvectors. This has been demonstrated by Meyer and Stewart [MS] in the context of complex parameters with $m = 1$ and the classical eigenvalue problem. In the context of (5.3), and especially (7.3), the solution is neatly expressed in terms of the group inverse of $L = L(\boldsymbol{\rho}_0, \lambda(\boldsymbol{\rho}_0))$, say $L^\#$, when the zero eigenvalue of the constant matrix L is semisimple.

There is a distinction here between the matrix function $L(\lambda)$ evaluated at λ_0 and the matrix function $L(\lambda_0) - \lambda I$. Let us illustrate the point with an example. Let $a_1(\lambda), a_2(\lambda)$ be polynomials for which $a_1(\lambda_0)a_2(\lambda_0) \ne 0$ and let

(8.1)
$$L(\lambda) = \begin{bmatrix} 0 & a_2(\lambda) \\ (\lambda - \lambda_0)a_1(\lambda) & 0 \end{bmatrix}.$$

Clearly, $L(\lambda)$ has a simple zero at λ_0. However,

$$L(\lambda_0) = \begin{bmatrix} 0 & a_2(\lambda_0) \\ 0 & 0 \end{bmatrix}$$

and has a double zero in $\det(L(\lambda_0)-\lambda I)$ with a nonlinear elementary divisor (and rank $L^2 \neq \text{rank}L$). In this case the use of the group inverse in the solution of (7.3) is *not* appropriate. Thus, when solving (7.3) in terms of $L^{\#}$ it will be necessary to make the further hypothesis that $\text{rank}L^2 = \text{rank}L$ at $\lambda_0 = \lambda(\rho_0)$. This is automatically satisfied if $L(\rho,\lambda)$ comes from a classical eigenvalue problem, when $L(\rho,\lambda) = A(\rho) - \lambda I$, and also when $L(\rho,\lambda)$ takes values in the hermitian matrices.

For the definition and properties of the group inverse, refer to the standard references [BG] and [CMe]. We would remark however that, for a singular matrix A with $\text{rank}A^2 = \text{rank}A$, the group inverse can be defined by the functional calculus (see [LT], for example). It is the function $A^{\#} = f(A)$ obtained by defining

$$f(\lambda) = \begin{cases} 1/\lambda & \text{when } \lambda \neq 0, \\ 0 & \text{when } \lambda = 0. \end{cases}$$

For this reason the term "spectral inverse" for group inverse is also appropriate.

The matrix $A^{\#}A$ is a projector (generally skew) onto ImA along ker$A$ and $A^{\#}A = AA^{\#}$. Note the strong contrast with properties of the Moore–Penrose inverse. When A^2 and A have the same rank the general solution of the consistent equation

$$A\mathbf{y} = \mathbf{b}$$

has the form

$$\mathbf{y} = A^{\#}\mathbf{b} + \alpha\mathbf{x},$$

where $\mathbf{x} \in \ker A$ and α is any complex number.

THEOREM 8.1. *Given the hypotheses of Theorem* 7.1 *and also* $\text{rank}L^2 = \text{rank}L$ *at* $(\rho_0, \lambda(\rho_0))$, *then the unique solution of* (5.3) *is given by* (7.2) *and*

$$\mathbf{x}_j = -(I - \mathbf{x}_0\mathbf{z}^T)L^{\#}(\lambda_j L^{(1)} + L_j)\mathbf{x}_0. \tag{8.2}$$

When $\mathbf{z} = \mathbf{y}_0$, *a left eigenvector of* L, *corresponding to the eigenvalue* λ_0,

$$\mathbf{x}_j = -L^{\#}(\lambda_j L^{(1)} + L_j)\mathbf{x}_0. \tag{8.3}$$

Proof. All solutions of (7.3) have the form

$$\mathbf{x}_j = -L^{\#}(\lambda_j L^{(1)} + L_j)\mathbf{x}_0 + \alpha\mathbf{x}_0$$

for some scalar α. Acting on this equation with the projector $I - \mathbf{x}_0\mathbf{z}^T$ (as in the proof of Theorem 7.1) gives (8.2) and this $\mathbf{x}_j$ obviously satisfies $\mathbf{z}^T\mathbf{x}_j = 0$. Thus (5.3) is satisfied.

By definition of a left eigenvector, $L^T\mathbf{y}_0 = \mathbf{0}$, and so $\mathbf{y}_0^T L = \mathbf{0}^T$. Since Im $L^{\#} =$ ImL we also have $\mathbf{y}_0^T L^{\#} = \mathbf{0}^T$. Thus the choice $\mathbf{z} = \mathbf{y}_0$ in (8.2) produces (8.3). □

Note that, when $\mathbf{z} = \mathbf{y}_0$, $\mathbf{x}_0\mathbf{y}_0^T$ and $I - \mathbf{x}_0\mathbf{y}_0^T$ are spectral projectors for L at ρ_0 (i.e., in the sense of the classical eigenvalue problem involving $L(\rho_0) - \lambda I$). See §4.10 of [LT], for example.

In contrast to §7 and especially Theorem 7.2, where $\| \mathbf{x}_j \|$ is estimated in terms of the singular values of L, the use of the group inverse admits similar estimates in terms of eigenvalues of L. Let $\mathbf{w}^T$ be any left eigenvector associated with a nonzero eigenvalue of L. Say $\mathbf{w}^T L = \mu\mathbf{w}^T$ where $\mu \neq 0$. Then $\mathbf{w}^T L^{\#} = \mu^{-1}\mathbf{w}^T$. We may assume $\| \mathbf{w} \| = 1$ and, when rank $L^2 = \text{rank}L$, $\mathbf{w}^T\mathbf{x}_0 = 0$.

Multiply (8.2) on the left by $\mathbf{w}^T$ and use these facts to obtain

$$\mathbf{w}^T\mathbf{x}_j = -\mu^{-1}\mathbf{w}^T(\lambda_j L^{(1)} + L_j)\mathbf{x}_0,$$

and hence

$$\frac{|\mathbf{w}^T(\lambda_j L^{(1)} + L_j)\mathbf{x}_0|}{|\mu|} \leq \| \mathbf{x}_j \| .$$

THEOREM 8.2. *Given the hypotheses of Theorem* 8.1, *let* $\mathbf{w}$ *be any left eigenvector for* L *such that* $\mathbf{w}^T L = \mu \mathbf{w}^T$, $\mu \neq 0$, *and* $\| \mathbf{w} \| = 1$. *Then*

$$\frac{|\mathbf{w}^T(\lambda_j L^{(1)} + L_j)\mathbf{x}_0|}{|\mu|} \leq \| \mathbf{x}_j \| \leq \| \mathbf{z} \| (1 + \| \mathbf{z} \|^{-1})^{1/2} \| L^{\#}(\lambda_j L^{(1)} + L_j)\mathbf{x}_0 \| . \tag{8.4}$$

Proof. The lower bound has just been established. The upper bound follows from (8.2) and (7.7). □

Note that when $\mathbf{z}^T = \mathbf{x}_0^*$ or when $\mathbf{z} = \mathbf{y}_0$, the terms in $\mathbf{z}$ on the right are replaced by "one."

As discussed in §7, in the neighbourhood of a multiple semisimple eigenvalue (with analytic extensions to simple eigenvalues) the unbounded growth of $\| \mathbf{x}_j \|$ suggested by (8.4) may be avoided as a result of orthogonality conditions following from the consistency of (5.2). Also, some sensitivity to numerical errors is generally unavoidable.

9. Higher derivatives with orthogonally constrained eigenvectors. In this section we maintain an important hypothesis of Theorems 7.1 and 8.1, namely, that the normalization chosen is the orthogonal constraint case of §4, i.e., that $\mathbf{z}$ in (5.3) is constant. Under the hypotheses of these two theorems, and noting the results of §6, it is clear that explicit formulae for higher derivatives can also be written down.

Thus under the hypotheses of Theorem 7.1, (6.1) has a unique solution given by

$$\lambda_{jk} = -\frac{\mathbf{y}_0^T \mathbf{b}_1}{\mathbf{y}_0^T L^{(1)} \mathbf{x}_0}, \tag{9.1}$$

where $\mathbf{y}_0$ is a left eigenvector of L at ρ_0, $\mathbf{b}_1$ is given in (6.2), and

$$\mathbf{x}_{jk} = -L^+(\lambda_{jk} L^{(1)} \mathbf{x}_0 + \mathbf{b}_1) + \alpha \mathbf{x}_0 \tag{9.2}$$

for some α. Now α must be chosen so that $\mathbf{z}^T \mathbf{x}_{jk} = 0$ (see (6.1) and (6.2)) and we conclude that

$$\mathbf{x}_{jk} = -(I - \mathbf{x}_0 \mathbf{z}^T) L^+(\lambda_{jk} L^{(1)} \mathbf{x}_0 + \mathbf{b}_1). \tag{9.3}$$

As in §7, this simplifies with the choice $\mathbf{z}^T = \mathbf{x}_0^*$ to

$$\mathbf{x}_{jk} = -L^+(\lambda_{jk} L^{(1)} \mathbf{x}_0 + \mathbf{b}_1), \tag{9.4}$$

the second derivative vector of minimal norm in the sense of Theorem 7.1.

Under the hypotheses of Theorem 8.1, $\mathbf{x}_{jk}$ can also be expressed in terms of the group inverse of L, of course.

Finally, note that, for the second derivative of a uniformly normed eigenvector, a different choice of α must be made in (9.2).

10. The spectral condition number and bordered matrices. A result of Chu [C2] concerning the spectral condition number of bordered matrices quantifies some earlier remarks on the effects of bordering a singular matrix, as in (5.3) and (6.1). The relevant result appears on page 699 of [C2] for real matrices and is adapted here to the complex case. First, as a measure of the angle between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$ we use θ, where

$$\cos\theta = \frac{|\mathbf{u}^* \mathbf{v}|}{\|\mathbf{u}\| \|\mathbf{v}\|}, \qquad 0 \leq \theta \leq \frac{\pi}{2}$$

and write $\cos\theta = \cos(\mathbf{u},\mathbf{v})$. The "spectral condition number" of a square matrix A is denoted by $\kappa(A) = \|A\|\|A^{-1}\|$. Let

$$\tilde{L} = \begin{bmatrix} L & \mathbf{q} \\ \mathbf{z}^T & 0 \end{bmatrix} \tag{10.1}$$

and, as in §7, let $L = UDV^*$ be a SVD of L, with D having the form (7.8) with $\sigma_1 \geq \cdots \geq \sigma_{n-1}$. A simpler argument than that in [C2] establishes the following result, which is also valid in the complex case and, unlike the result in [C2], does not assume $\| \mathbf{z} \| = \| \mathbf{q} \| = 1$.

THEOREM 10.1. *With the above notations and conventions, suppose, in addition, that* $\sigma_1 \geq 1 \geq \sigma_{n-1}$. *Let* λ_0 *be a simple eigenvalue of* L *with right and left eigenvectors* $\mathbf{x},\mathbf{y}$, *respectively. Then if* $\mathbf{z}^T\mathbf{x} \neq 0$ *and* $\mathbf{y}^T\mathbf{q} \neq 0$, *we have*

$$\frac{\sigma_1}{\sigma_{n-1}} \leq \kappa(\tilde{L}) \leq \frac{(1+ \| \mathbf{q} \| + \| \mathbf{z} \|)(2+ \| \mathbf{q} \|^{-1} + \| \mathbf{z} \|^{-1})}{\cos(\bar{\mathbf{z}},\mathbf{x})\cos(\mathbf{q},\bar{\mathbf{y}})} \frac{\sigma_1}{\sigma_{n-1}}. \tag{10.2}$$

Proof. The SVD of L may be written

$$L = [U_1, \mathbf{u}_n] \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0}^T & 0 \end{bmatrix} \begin{bmatrix} V_1^* \\ \mathbf{v}_n^* \end{bmatrix},$$

where $\mathbf{u}_n$ and $\mathbf{v}_n$ have the same direction as $\bar{\mathbf{y}}$ and $\mathbf{x}$, respectively. Then

$$\tilde{L} = \begin{bmatrix} U_1 & \mathbf{u}_n & \mathbf{0} \\ \mathbf{0}^T & 0 & 1 \end{bmatrix} \begin{bmatrix} \Sigma & \mathbf{0} & \mathbf{d}_1 \\ \mathbf{0}^T & 0 & d_2 \\ \mathbf{c}_1^* & \bar{c}_2 & 0 \end{bmatrix} \begin{bmatrix} V_1^* & \mathbf{0} \\ \mathbf{v}_n^* & 0 \\ \mathbf{0}^T & 1 \end{bmatrix},$$

where

$$\mathbf{q} = [U_1, \mathbf{u}_n] \begin{bmatrix} \mathbf{d}_1 \\ d_2 \end{bmatrix} = U\mathbf{d} \quad \text{and} \quad \bar{\mathbf{z}} = [V_1, \mathbf{v}_n] \begin{bmatrix} \mathbf{c}_1 \\ c_2 \end{bmatrix} = V\mathbf{c}.$$

Also, $\mathbf{z}^T\mathbf{x} \neq 0$ and $\mathbf{y}^T\mathbf{q} \neq 0$ imply $d_2 \neq 0$ and $c_2 \neq 0$. Hence, if $\cos\theta = \cos(\mathbf{q},\mathbf{u}_n) = \cos(\mathbf{q},\bar{\mathbf{y}})$ and $\cos\phi = \cos(\bar{\mathbf{z}},\mathbf{v}_n) = \cos(\bar{\mathbf{z}},\mathbf{x})$, then

$$\cos\theta = \frac{|d_2|}{\| \mathbf{q} \|}, \quad \sin\theta = \frac{\| \mathbf{d}_1 \|}{\| \mathbf{q} \|}, \quad \tan\theta = \| d_2^{-1}\mathbf{d}_1 \|,$$
$$\cos\phi = \frac{|c_2|}{\| \mathbf{z} \|}, \quad \sin\phi = \frac{\| \mathbf{c}_1 \|}{\| \mathbf{z} \|}, \quad \tan\phi = \| c_2^{-1}\mathbf{c}_1 \|,$$

and

$$\tilde{L}^{-1} = \begin{bmatrix} V_1 & \mathbf{v}_n & \mathbf{0} \\ \mathbf{0}^T & 0 & 1 \end{bmatrix} \begin{bmatrix} \Sigma^{-1} & -d_2^{-1}\Sigma^{-1}\mathbf{d}_1 & \mathbf{0} \\ -\bar{c}_2^{-1}\mathbf{c}_1^*\Sigma^{-1} & \bar{c}_2^{-1}d_2^{-1}\mathbf{c}_1^*\Sigma^{-1}\mathbf{d}_1 & \bar{c}_2^{-1} \\ \mathbf{0}^T & d_2^{-1} & 0 \end{bmatrix} \begin{bmatrix} U_1^* & \mathbf{0} \\ \mathbf{u}_n^* & 0 \\ \mathbf{0}^T & 1 \end{bmatrix}.$$

It is easily verified that $\| \tilde{L} \| \leq \| \Sigma \| + \| \mathbf{d} \| + \| \mathbf{c} \| = \sigma_1 + \| \mathbf{q} \| + \| \mathbf{z} \|$, and hence

$$\sigma_1 \leq \sigma_{\max}(\tilde{L}) \leq \sigma_1 + \| \mathbf{q} \| + \| \mathbf{z} \| \leq \sigma_1(1+ \| \mathbf{q} \| + \| \mathbf{z} \|).$$

Similarly, it is found that

$$\sigma_{n-1}^{-1} \leq \sigma_{\min}^{-1}(\tilde{L}) \leq \sigma_{n-1}^{-1}\left\{1+\tan\theta+\tan\phi+\tan\theta\tan\phi+\frac{1}{\| \mathbf{q} \| \cos\theta}+\frac{1}{\| \mathbf{z} \| \cos\phi}\right\}$$

$$\leq \frac{\sigma_{n-1}^{-1}}{\cos\theta\cos\phi}\{2+ \| \mathbf{q} \|^{-1} + \| \mathbf{z} \|^{-1}\}.$$

The result follows. □

A more elaborate argument, similar to that in [C2], yields a sharper upper bound in (10.2) which, when $\| \mathbf{z} \| = \| \mathbf{q} \| = 1$, reduces to

$$\frac{2\sqrt{5}}{\cos(\bar{\mathbf{z}},\mathbf{x})\cos(\mathbf{q},\bar{\mathbf{y}})} \quad \frac{\sigma_1}{\sigma_{n-1}}.$$

(The square root in the denominator of the corresponding result in [C2, p. 699] was included there by mistake and should be deleted, and the cosines in inequality (31) of [C2] should be squared.)

Note that the hypothesis $\sigma_1 \geq 1 \geq \sigma_{n-1}$ is not restrictive and can be achieved by scaling $L(\lambda)$. Theorem 10.1 can be interpreted as follows: the bordered matrix $\tilde{L}$ will be ill conditioned ($\kappa(\tilde{L})$ is "large") only if at least one of four conditions holds:

(i) σ_{n-1} is small compared to σ_1 ($\tilde{L}$ is singular if $\sigma_{n-1} = 0$);
(ii) $\bar{\mathbf{z}}$ is deficient in $\mathbf{x}$ ($\tilde{L}$ is singular if $\bar{\mathbf{z}} \perp \mathbf{x}$);
(iii) $\bar{\mathbf{q}}$ is deficient in $\mathbf{y}$ ($\tilde{L}$ is singular if $\bar{\mathbf{q}} \perp \mathbf{y}$); or
(iv) $\mathbf{q}$ or $\mathbf{z}$ is too large or too small.

Conversely, $\tilde{L}$ is ill conditioned if (i) holds. As (iv) can be avoided by standard prescaling techniques and σ_1/σ_{n-1} is a condition number for the singular system $L\mathbf{t} = \mathbf{g}$, the bounds in (10.2) show that the bordered matrix problem (suitably scaled) has similar conditioning to that of $L\mathbf{t} = \mathbf{g}$ unless (ii) or (iii) hold.

For our problems we have $\mathbf{q} = L^{(1)}\mathbf{x}$, so the freedom to align $\mathbf{q}$ with $\mathbf{y}$ is not available. However, Theorem 3.2 shows that, for simple eigenvalues, $\mathbf{y}^T L^{(1)}\mathbf{x} \neq 0$ so that, at least, $\cos(L^{(1)}\mathbf{x}, \bar{\mathbf{y}}) \neq 0$ (see also the discussion of §§3 and 4).

Of course, Theorem 10.1 concerns only the coefficient matrix of (5.3) and cannot address the point raised in §§7 and 8, that bounded solutions of (5.3) *may* exist in the limit as $\sigma_{n-1} \to 0$.

The ratio σ_1/σ_{n-1} also plays a role in the formulation of a "condition number" for the derivative vector $\mathbf{x}_j$. For simplicity, consider the case $\mathbf{z}^T = \mathbf{x}_0^*$ of Theorem 7.1. Substitute the expression (7.2) for λ_j into (7.6) to obtain

$$\mathbf{x}_j = -L^+ P L_j \mathbf{x}_0, \tag{10.3}$$

where

$$P = I - \left(\frac{1}{\mathbf{y}_0^T L^{(1)}\mathbf{x}_0}\right) L^{(1)}\mathbf{x}_0\mathbf{y}_0^T,$$

and $P^2 = P$ (i.e., P is the projector into $\{\text{span } (\bar{\mathbf{y}}_0)\}^\perp$ along $\text{span}(L^{(1)}\mathbf{x}_0)$). Standard techniques show that

$$\| P \| = \frac{\| L^{(1)}\mathbf{x}_0 \| \| \mathbf{y}_0 \|}{|\mathbf{y}_0^T L^{(1)}\mathbf{x}_0|} = \sec\,(\bar{\mathbf{y}}_0, L^{(1)}\mathbf{x}_0).$$

Since $\| L \| = \sigma_1$ and $\| L^+ \| = \sigma_{n-1}^{-1}$ (see [LT, Prop. 12.9.4]), it follows from (10.3) that, under the condition of Theorem 7.1,

$$\frac{\| \mathbf{x}_j \|}{\| \mathbf{x}_0 \|} \le \left(\frac{\sigma_1}{\sigma_{n-1}} \right) \left(\frac{\| L_j \|}{\| L \|} \right) \sec (\bar{\mathbf{y}}_0, L^{(1)}\mathbf{x}_0).$$

This bound demonstrates the possible sensitivity of $\| \mathbf{x}_j \|$ to three factors: the condition of L (reflecting closeness of λ_0 to another eigenvalue), the size of L_j, and the angle between $\bar{y}_0$ and $L^{(1)}\mathbf{x}_0$, respectively. Results of this kind for the classical eigenvalue problem are well known (see [W, p. 69], and also [GV, §7.2.4]).

11. Hermitian matrix functions. When the function $L(\rho, \lambda)$ takes values in the hermitian matrices then, generally, $\rho \in \mathbb{R}^m$. However, when the dependence on λ is nonlinear the eigenvalues may be complex. Furthermore, when L is hermitian-valued, but not real symmetric, then the eigenvector will be complex. These phenomena arise even in the quadratic eigenvalue problems (see (1.2)). In vibration theory the study of viscously damped systems leads to real-symmetric matrix functions, but if gyroscopic forces are admitted then complex hermitian matrices must be considered (see [BL], [L], or [Ro] for example).

Efficient numerical algorithms for the solution of (5.4) would then take advantage of the symmetry of L. From the analytical point of view there is, however, not very much to add to the discussion of explicit solutions for (5.4). The most notable feature is that, when L is hermitian, the group inverse always exists and is identical with the Moore–Penrose inverse. In contrast to the discussion of §8, a simple eigenvalue λ_0 of the matrix function determines a simple zero eigenvalue (in the classical sense) of the constant matrix $L(\lambda_0)$.

12. The classical eigenvalue problem. Although the main thrust of this paper is in the direction of problems with nonlinear dependence on λ, it may be useful to present some conclusions for the special case of the classical eigenvalue problem, i.e., when

$$L(\rho, \lambda) = A(\rho) - \lambda I \tag{12.1}$$

and the eigenvalue is simple. The first observation is, of course, that explicit formulae of §§5–9 simplify in view of the fact that

$$L^{(1)} = -I, \; L^{(2)} = L^{(3)} = \cdots = 0.$$

Furthermore, $L_j^{(1)} = 0$ for $j = 1, 2, \ldots, m$. Let us first consider an *orthogonally constrained eigenvector*, i.e., for which $\|\mathbf{x}(\rho_0)\| = 1$, $\mathbf{z}^T\mathbf{x}(\rho) = 1$ and $\mathbf{z}$ is independent of ρ. This normalization determines derivative vectors of $\mathbf{x}(\rho)$ whose norm is minimized by the choice $\mathbf{z}^T = \mathbf{x}_0^*$.

The equations

$$\begin{bmatrix} A - \lambda I & -\mathbf{x}_0 \\ \mathbf{z}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_j \\ \lambda_j \end{bmatrix} = - \begin{bmatrix} A_j\mathbf{x}_0 \\ 0 \end{bmatrix} \tag{12.2}$$

and

$$\begin{bmatrix} A - \lambda I & -\mathbf{x}_0 \\ \mathbf{z}^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_{jk} \\ \lambda_{jk} \end{bmatrix} = - \begin{bmatrix} (A_k - \lambda_k I)\mathbf{x}_j + (A_j - \lambda_j I)\mathbf{x}_k + A_{jk}\mathbf{x}_0 \\ 0 \end{bmatrix} \tag{12.3}$$

have the unique solution

$$\lambda_j = \mathbf{y}_0^T A_j \mathbf{x}_0, \tag{12.4}$$

$$\mathbf{x}_j = -(I - \mathbf{x}_0\mathbf{z}^T)(A - \lambda I)^+(A_j - \lambda_j I)\mathbf{x}_0, \tag{12.5}$$

$$\lambda_{jk} = \mathbf{y}_0^T\{(A_k - \lambda_k I)\mathbf{x}_j + (A_j - \lambda_j I)\mathbf{x}_k + A_{jk}\mathbf{x}_0\}, \tag{12.6}$$

$$\mathbf{x}_{jk} = -(I - \mathbf{x}_0\mathbf{z}^T)(A - \lambda I)^+\{(A_k - \lambda_k I)\mathbf{x}_j + (A_j - \lambda_j I)\mathbf{x}_k + A_{jk}\mathbf{x}_0 - \lambda_{jk}\mathbf{x}_0\}, \tag{12.7}$$

where $\mathbf{y}_0$ is a left eigenvector at λ for which $\mathbf{y}_0^T\mathbf{x}_0 = 1$.

Two ways in which these formulae simplify are noteworthy. First, the group inverse $(A - \lambda I)^{\#}$ can be used instead of the Moore–Penrose inverse (see §8). Then we have $(A - \lambda I)^{\#}\mathbf{x}_0 = \mathbf{0}$ and the formulae for $\mathbf{x}_j, \mathbf{x}_{jk}$ simplify accordingly. Furthermore, if we then choose $\mathbf{z} = \mathbf{y}_0$, then $\mathbf{y}_0^T(A - \lambda I)^{\#} = \mathbf{0}^T$. Thus

$$\mathbf{x}_j = -(A - \lambda I)^{\#}A_j\mathbf{x}_0 \tag{12.8}$$

$$\mathbf{x}_{jk} = -(A - \lambda I)^{\#}\{(A_k - \lambda_k I)\mathbf{x}_j + (A_j - \lambda_j I)\mathbf{x}_k + A_{jk}\mathbf{x}_0\}. \tag{12.9}$$

Second, (12.5) and (12.7) simplify if we choose $\mathbf{z}^T = \mathbf{x}_0^*$, because in this case $\mathbf{z}^T(A - \lambda I)^+ = \mathbf{x}_0^*(A - \lambda I)^+ = \mathbf{0}^T$.

Formulae involving the group inverse are consistent with results from Theorem 1 in the paper of Meyer and Stewart [MS].

Uniformly normed eigenvector. In this case the equations determining first and second derivatives are

$$\begin{bmatrix} A - \lambda I & -\mathbf{x}_0 \\ \mathbf{x}_0^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_j \\ \lambda_j \end{bmatrix} = -\begin{bmatrix} A_j\mathbf{x}_0 \\ 0 \end{bmatrix}, \tag{12.10}$$

$$\begin{bmatrix} A - \lambda I & -\mathbf{x}_0 \\ \mathbf{x}_0^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x}_{jk} \\ \lambda_{jk} \end{bmatrix} = -\begin{bmatrix} (A_k - \lambda_k I)\mathbf{x}_j + (A_j - \lambda_j I)\mathbf{x}_k + A_{jk}\mathbf{x}_0 \\ \frac{1}{2}(\mathbf{x}_j^*\mathbf{x}_k + \mathbf{x}_k^*\mathbf{x}_j) \end{bmatrix}. \tag{12.11}$$

Formulae analogous to (12.4)–(12.7) are:

$$\lambda_j = \mathbf{y}_0^T A_j \mathbf{x}_0, \tag{12.12}$$

$$\mathbf{x}_j = -(A - \lambda I)^+(A_j - \lambda_j I)\mathbf{x}_0, \tag{12.13}$$

$$\lambda_{jk} = \mathbf{y}_0^T\{(A_k - \lambda_k I)\mathbf{x}_j + (A_j - \lambda_j I)\mathbf{x}_k + A_{jk}\mathbf{x}_0\}, \tag{12.14}$$

$$\mathbf{x}_{jk} = -(A - \lambda I)^+\{(A_k - \lambda_k I)\mathbf{x}_j + (A_j - \lambda_j I)\mathbf{x}_k + A_{jk}\mathbf{x}_0 - \lambda_{jk}\mathbf{x}_0\} - \tfrac{1}{2}(\mathbf{x}_j^*\mathbf{x}_k + \mathbf{x}_k^*\mathbf{x}_j)\mathbf{x}_0. \tag{12.15}$$

Corresponding formulae using the group inverse are left for the interested reader.

Classical hermitian problems. Finally, let us summarize the derivative formulae for hermitian matrix functions when the eigenvalue is simple and, necessarily, ρ takes only real values. Note that the Moore–Penrose and group inverses now coincide.

For an orthogonally constrained eigenvector:

$$\begin{aligned}
\lambda_j &= \mathbf{x}_0^* A_j \mathbf{x}_0, \\
\mathbf{x}_j &= -(I - \mathbf{x}_0\mathbf{z}^T)(A - \lambda I)^+ A_j \mathbf{x}_0, \\
\lambda_{jk} &= \mathbf{x}_0^*(A_k\mathbf{x}_j + A_j\mathbf{x}_k + A_{jk}\mathbf{x}_0), \\
\mathbf{x}_{jk} &= -(I - \mathbf{x}_0\mathbf{z}_0^T)(A - \lambda I)^+\{(A_k - \lambda_k I)\mathbf{x}_j + (A_j - \lambda_j I)\mathbf{x}_k + A_{jk}\mathbf{x}_0\}.
\end{aligned}$$

For a uniformly normed eigenvector:

$$\begin{aligned}
\lambda_j &= \mathbf{x}_0^* A_j \mathbf{x}_0, \\
\mathbf{x}_j &= -(A-\lambda I)^+ A_j \mathbf{x}_0, \\
\lambda_{jk} &= \mathbf{x}_0^*(A_k \mathbf{x}_j + A_j \mathbf{x}_k + A_{jk}\mathbf{x}_0), \\
\mathbf{x}_{jk} &= -(A-\lambda I)^+\{(A_k-\lambda_k I)\mathbf{x}_j + (A_j - \lambda_j I)\mathbf{x}_k + A_{jk}\mathbf{x}_0\} - \tfrac{1}{2}(\mathbf{x}_j^*\mathbf{x}_k + \mathbf{x}_k^*\mathbf{x}_j)\mathbf{x}_0.
\end{aligned}$$

Appendix. Bordered matrices and continuation methods. Linear systems with bordered coefficient matrix, as in (5.4), arise naturally in the study of "fully nonlinear" problems. This connection is clarified here using the methods of Rheinboldt [Rh]. Similar results can be developed using the techniques of Keller et al. (see [Ke2, p. 71] or [MoS, p. 137]).

Consider first (5.1) where $\mathbf{z}(\boldsymbol{\rho})$ is a given analytic function and define a map

$$G : \mathbb{C}^n \times \mathcal{F} \times \mathcal{F}^m \to \mathbb{C}^n \times \mathbb{C}$$

by

$$G(\mathbf{x},\lambda,\boldsymbol{\rho}) = \begin{cases} L(\boldsymbol{\rho},\lambda)\mathbf{x}, \\ \mathbf{z}^T\mathbf{x} - 1. \end{cases}$$

Here, $\mathcal{F}$ denotes $\mathbb{R}$ or $\mathbb{C}$ according as λ and $\boldsymbol{\rho}$ take real or complex values.

Denote the $(n+1)\times(n+m+1)$ Jacobian matrix of G by ΔG. If ΔG has full rank at $(\mathbf{x},\lambda,\boldsymbol{\rho})$ this point is said to be *regular*, otherwise $(\mathbf{x},\lambda,\boldsymbol{\rho})$ is a *singular* point. We have

$$\Delta G = \begin{bmatrix} L & L^{(1)}\mathbf{x} & L_1\mathbf{x} & \dots & L_m\mathbf{x} \\ \mathbf{z}^T & 0 & \mathbf{z}_1^T\mathbf{x} & \dots & \mathbf{z}_m^T\mathbf{x} \end{bmatrix}.$$

It has been seen (as in (5.3)) that analytic dependence of $\lambda, \mathbf{x}$, and $\mathbf{z}$, on $\boldsymbol{\rho}$ imply that, for $j = 1,2,\dots,m$

$$\begin{bmatrix} L_j\mathbf{x} \\ \mathbf{z}_j^T\mathbf{x} \end{bmatrix} \in \operatorname{Im}\begin{bmatrix} L & L^{(1)}\mathbf{x} \\ \mathbf{z}^T & 0 \end{bmatrix},$$

so in this case

$$\operatorname{rank}(\Delta G) = \operatorname{rank}\begin{bmatrix} L & L^{(1)}\mathbf{x} \\ \mathbf{z}^T & 0 \end{bmatrix}.$$

Thus, under the conditions of Corollary 5.5 (in particular, when λ_0 is a simple eigenvalue), $(\mathbf{x},\lambda_0,\boldsymbol{\rho})$ is a regular point of G. If, however, λ_0 is semisimple with multiplicity two or more, then $(\mathbf{x},\lambda_0,\boldsymbol{\rho})$ is a singular point of G.

In the context of uniformly normed eigenvectors, we define

$$F : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^n \times \mathbb{R}$$

by

$$F(\mathbf{x},\lambda,\boldsymbol{\rho}) = \begin{cases} L(\boldsymbol{\rho},\lambda)\mathbf{x}, \\ \mathbf{x}^T\mathbf{x} - 1 \end{cases}$$

and assume, for the sake of the existence of necessary derivatives, that all parameters are real.

Now it is found that

$$\Delta F = \begin{bmatrix} L & L^{(1)}\mathbf{x} & L_1\mathbf{x} & \dots & L_m\mathbf{x} \\ 2\mathbf{x}^T & 0 & \alpha_1 & \dots & \alpha_m \end{bmatrix},$$

where $\alpha_j = 2\mathbf{x}_j^T\mathbf{x}, j = 1, \ldots, m$. But then (4.7) shows that $\alpha_1 = \cdots = \alpha_m = 0$, and it follows from (5.2) that

$$\operatorname{rank}(\Delta F) = \operatorname{rank}\begin{bmatrix} L & L^{(1)}\mathbf{x} \\ 2\mathbf{x}^T & 0 \end{bmatrix}.$$

Again, if λ_0 is semisimple, $(\mathbf{x}, \lambda_0, \boldsymbol{\rho})$ is a regular or singular point when λ_0 has multiplicity one, or more, respectively.

Acknowledgment. The third-named author is grateful to Professor E. Bohl for the opportunity to present some of this work in his seminar at the University of Konstanz.

REFERENCES

[A] A. L. ANDREW, *Computation of higher Sturm-Liouville eigenvalues*, Congr. Numer., 34 (1982), pp. 3–16.

[ACL] A. L. ANDREW, K. W. E. CHU, AND P. LANCASTER, *Sensitivities of eigenvalues and eigenvectors of problems nonlinear in the eigenparameter*, Appl. Math. Lett., 5 (1992), pp. 69–72.

[AHT] A. L. ANDREW, F. R. DE HOOG, AND R. C. E. TAN, *Direct computation of derivatives of eigenvalues and eigenvectors*, Internat. J. Comput. Math., 36 (1990), pp. 251–255.

[BL] L. BARKWELL AND P. LANCASTER, *Overdamped and gyroscopic vibrating systems*, Trans. ASME J. Appl. Mech., 59 (1992), pp. 176–181.

[B] H. BAUMGÄRTEL, *Analytic Perturbation Theory for Matrices and Operators*, Birkhäuser Verlag, Basel, 1985.

[BG] A. BEN-ISRAEL AND T. N. E. GREVILLE, *Generalized Inverses: Theory and Applications*, John Wiley, New York, 1974.

[BM] S. BOCHNER AND W. T. MARTIN, *Several Complex Variables*, Princeton University Press, Princeton, NJ, 1948.

[CMe] S. L. CAMPBELL AND C. D. MEYER, JR., *Generalized Inverses of Linear Transformations*, Pitman, London, 1979.

[CMa] C. CARDANI AND P. MANTEGAZZA, *Calculation of eigenvalue and eigenvector derivatives for algebraic flutter and divergence eigenproblems*, AIAA J., 17 (1979), pp. 408–412.

[Ch] T. F. CHAN, *An approximate Newton method for coupled nonlinear systems*, SIAM J. Numer. Anal., 22 (1985), pp. 904–913.

[CR] T. F. CHAN AND D. C. RESASCO, *On the condition of nearly singular matrices under rank*-1 *perturbation*, Linear Algebra Appl., 76 (1986), pp. 223–232.

[C1] K.-W. E. CHU, *On multiple eigenvalues of matrices depending on several parameters*, SIAM J. Numer. Anal., 27 (1990), pp. 1368–1385.

[C2] ———, *Bordered matrices, singular systems and ergodic Markov chains*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 688–701.

[Ga] S. GARG, *Derivatives of eigensolutions for a general matrix*, AIAA J., 11 (1973), pp. 1191–1194.

[GLR] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[GR] I. GOHBERG AND L. RODMAN, *Analytic matrix functions with prescribed local data*, J. Analyse Math., 40 (1981), pp. 90–128.

[GV] G. H. GOLUB AND C. VAN LOAN, *Matrix computations*, North Oxford Academic, Oxford, 1984.

[HA] R. T. HAFTKA AND H. M. ADELMAN, *Recent developments in structural sensitivity analysis*, Structural Optimization, 1 (1989), pp. 137–151.

[HC] E. J. HAUG AND K. K. CHOI, *Systematic occurrence of repeated eigenvalues in structural optimization*, J. Optim. Theory Appl., 38 (1982), pp. 251–274.

[HR] E. J. HAUG AND B. ROUSSELET, *Design sensitivity analysis in structural mechanics* II. *Eigenvalue variations*, J. Structural Mech., 8 (1980), pp. 161–186.

[Ka] T. KATO, *Perturbation Theory for Linear Operators*, Springer Verlag, New York, 1966.

[Ke1] H. B. KELLER, *The bordering algorithm and path following near singular points of higher nullity*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 573–582.

[Ke2] ———, *Lectures on Numerical Methods in Bifurcation Problems*, Springer Verlag, New York, 1987.

[KK] V. B. KHAZANOV AND V. N. KUBLANOVSKAYA, *Spectral problems for matrix pencils. Methods and algorithms.* II, Soviet J. Numer. Anal. Math. Modelling, 3 (1988), pp. 467–485.

[L] P. LANCASTER, *Lambda-Matrices and Vibrating Systems*, Pergamon, Oxford, 1966.

[LT] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, 2nd ed., Academic Press, New York, 1985.

[MS] C. D. MEYER AND G. W. STEWART, *Derivatives and perturbations of eigenvectors*, SIAM J. Numer. Anal., 25 (1988), pp. 679–691.

[MoS] G. MOORE AND A. SPENCE, *The convergence of approximations to nonlinear equations at single turning points*, in Bifurcation Problems and Their Numerical Solution, H. D. Mittlemann and H. Weber, eds., Birkhäuser Verlag, Basel, 1980, pp. 135–146.

[Oj] I. U. OJALVO, *Efficient computation of mode-shape derivatives for large dynamic systems*, AIAA J., 25 (1987), pp. 1386–1390.

[OM] M. R. OSBORNE AND S. MICHAELSON, *The numerical solution of eigenvalue problems in which the eigenvalue appears nonlinearly, with an application to differential equations*, Computer J., 7 (1964), pp. 66–71.

[Rh] W. C. RHEINBOLDT, *Numerical Analysis of Parametrized Nonlinear Equations*, John Wiley, New York, 1986.

[Ro] M. ROSEAU, *Vibrations in Mechanical Systems*, Springer Verlag, Berlin, 1987.

[Ru] A. RUHE, *Algorithms for the nonlinear eigenvalue problem*, SIAM J. Numer. Anal., 10 (1973), pp. 674–689.

[S1] J. G. SUN, *Eigenvalues and eigenvectors of a matrix dependent on several parameters*, J. Comput. Math., 3 (1985), pp. 351–364.

[S2] ———, *Multiple eigenvalue sensitivity anaylysis*, Linear Algebra Appl., 137/138 (1990), pp. 183–211.

[T] R. C. E. TAN, *An extrapolation method for computing derivatives of eigensystems*, Internat. J. Comput. Math., 22 (1987), pp. 63–73.

[TA] R. C. E. TAN AND A. L. ANDREW, *Computing derivatives of eigenvalues and eigenvectors by simultaneous iteration*, IMA J. Numer. Anal., 9 (1989), pp. 111–122.

[W] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 927–949, October 1993

002

NUMERICAL METHODS FOR SIMULTANEOUS DIAGONALIZATION*

ANGELIKA BUNSE-GERSTNER[†], RALPH BYERS[‡], AND VOLKER MEHRMANN[§]

Abstract. A Jacobi-like algorithm for simultaneous diagonalization of commuting pairs of complex normal matrices by unitary similarity transformations is presented. The algorithm uses a sequence of similarity transformations by elementary complex rotations to drive the off-diagonal entries to zero. Its asymptotic convergence rate is shown to be quadratic and numerically stable. It preserves the special structure of real matrices, quaternion matrices, and real symmetric matrices.

Key words. simultaneous diagonalization, Jacobi iteration, eigenvalues, eigenvectors, structured eigenvalue problem

AMS subject classifications. 65F15, 65-04

1. Introduction. Many of the algorithms outlined in [6] require the simultaneous diagonalization of commuting pairs of normal matrices by unitary similarity transformations. Often there are other structures in addition to normality. Examples include commuting pairs of real symmetric matrices, pairs of Hermitian matrices, real symmetric—real skew symmetric pairs, and quaternion pairs. In this paper we point out some of the difficulties associated with simultaneous diagonalization and propose a family of Jacobi-like algorithms for simultaneously diagonalizing commuting normal matrices.

The term "simultaneous diagonalization" is sometimes used in the literature to denote the diagonalization of a definite matrix pencil by congruence, e.g., [43]. Here, we use the term in the classical sense of simultaneous similarity transformations.

To be viable for finite-precision computation, a simultaneous diagonalization algorithm must work with both A and B simultaneously. To see how algorithms that violate this principle can fail, consider the family of *diagonalize-one-then-diagonalize-the-other* (DODO) methods suggested by the classic proof that commuting pairs of diagonalizable matrices can be simultaneously diagonalized [30, p. 404]. If $A \in \mathbf{C}^{n\times n}$ and $B \in \mathbf{C}^{n\times n}$ form a pair of commuting normal matrices, the DODO approach uses a conventional algorithm to diagonalize A alone, and it then performs the same similarity transformation on B. So, for example, if A and B are Hermitian, then one might use `CH` from [39] to find a unitary matrix $U \in \mathbf{C}^{n\times n}$ and a diagonal matrix $D \in \mathbf{R}^{n\times n}$ such that $A = U^H DU$. (The superscript H denotes the Hermitian transpose.) Although `CH` does not produce the diagonal entries of D in any particular order, it is easy to order the eigenvalues in decreasing algebraic order along the diagonal of D. Then $E := UBU^T$ is block diagonal with the order of the jth diagonal block equal to the multiplicity of the jth distinct eigenvalue of A. In particular, if A has distinct

* Received by the editors March 14, 1990; accepted for publication (in revised form) January 17, 1992. The work of all three authors was partially supported by Forschungsschwerpunkt Mathematisierung Universität Bielefeld through research project Steuerungsprobleme.

[†] Fakultät für Mathematik, Universität Bielefeld, Postfach 8640, D-4800 Bielefeld 1, Germany (`umatf112@dbiuni11.bitnet`).

[‡] Department of Mathematics, University of Kansas, Lawrence, Kansas 66045 (`byers@kuhub.cc.ukans.edu`). The work of this author was partially supported by National Science Foundation grant CCR-8820882, and by the University of Kansas through International Travel Fund grant 560478, General Research Allocations grants 3758-20-0038 and 3692-20-0038.

[§] Fakultät für Mathematik, Universität Bielefeld, Postfach 8640, D-4800 Bielefeld 1, Germany (`umatf108@dbiuni11.bitnet`).

eigenvalues, then E is diagonal and the simultaneous diagonalization is complete. In any case, a subsequent block diagonal similarity transformation can diagonalize E without disturbing D.

Rounding errors destroy this elegant approach. Suppose, for example, rounding errors perturb the commuting pair (A, B) to the nearly commuting pair

$$\tilde{A} = \begin{bmatrix} 1-\epsilon & 0 & 0 & 0 \\ 0 & 1+\epsilon & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

and

$$\tilde{B} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1-\epsilon & 0 \\ 0 & 0 & 0 & 1+\epsilon \end{bmatrix},$$

where ϵ is a small quantity that might be caused by rounding error. If $\epsilon \neq 0$, then $\tilde{A}$ and $\tilde{B}$ do not commute. However, if

$$Q = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix},$$

then the off-diagonal elements of $Q^T \tilde{A} Q$ and $Q^T \tilde{B} Q$ are bounded by ϵ.

For $\epsilon \neq 0$, ± 2, $\tilde{A}$ has distinct eigenvalues, and its modal matrix of eigenvectors

$$U = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

is independent of ϵ and unique up to column scaling. Similarly, for $\epsilon \neq 0$, ± 2, B has distinct eigenvalues and its modal matrix of eigenvectors

$$V = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

is independent of ϵ and unique up to column scaling. Unfortunately, the off-diagonal entries of $U^T B U$ and $V^T A V$ are of the order of 1. The DODO method creates a cyclic sequence with period two.

The example suggests that a viable simultaneous diagonalization algorithm must "do something reasonable" when it is applied to a nearly commuting pair of matrices. Perhaps the most natural approach is to require the algorithm to choose a unitary similarity transformation $U \in \mathbf{C}^{n \times n}$ that minimizes some measure of how much $U^H A U$ and $U^H B U$ differ from being diagonal. This is the approach used in [21]. There, the algebraic eigenvalue problem of a single normal matrix is solved by simultaneously diagonalizing the Hermitian and skew-Hermitian parts. It is known to converge locally quadratically under the serial pivot sequence [36]. When applied to a normal matrix, the norm-reducing method of [14] is also an algorithm that simultaneously diagonalizes the Hermitian and skew-Hermitian part of a normal matrix. A variation of this method that uses nonunitary norm-reducing transformations is shown to be globally quadratically convergent for complex normal matrices in [17]. The simultaneous diagonalization algorithms presented below are adaptations and generalizations of [21] and are influenced by [14].

Interest in Jacobi algorithms declined when it was observed that classic and serial Jacobi algorithms for the symmetric eigenvalue problem perform more arithmetic operations than do more modern techniques [22], [33], [47]. However, with the advent of parallel and vector computers, interest in Jacobi methods has revived. Jacobi methods have high inherent parallelism, which allows efficient implementations on certain parallel architectures [4], [5], [8], [9], [13], [15], [16], [28], [27], [37]. This has been demonstrated on parallel and vector machines in [2], [3], [15]. Parallel orderings, block Jacobi methods, and other techniques create parallel versions of Jacobi's method. It is easy to see that these techniques also apply to the simultaneous diagonalization algorithms presented here. A short discussion of some special techniques for parallel computation along with a more extensive bibliography can be found in [22, §8.5].

Another virtue of the Jacobi method is its favorable rounding-error properties. Improved error bounds for perturbed scaled diagonally dominant matrices [1], [10], [25] show that for this class of matrices, small relative perturbations in the matrix entries cause only small relative perturbations in the eigenvalues. In some cases, Jacobi's method has rounding-error properties better than those of the QR algorithm [11]. We conjecture that this very favorable error analysis carries over to the simultaneous diagonalization process as well.

We shall use the following notation.

- The transpose of a matrix is denoted by a superscript T.
- The Hermitian or complex conjugate transpose of a matrix is denoted by a superscript H.
- The vector 2-norm and its subordinate matrix norm, i.e., the spectral norm, are denoted by $\|\cdot\|_2$.
- The Frobenius norm or Euclidean norm is denoted by $\|M\|_F = \sqrt{\text{trace}(M^H M)}$.
- The kth column of the identity matrix is denoted by e_k.
- The smallest singular value of a matrix $M \in \mathbf{R}^{n\times n}$ is denoted by $\sigma_{\min}(M)$.

2. A Jacobi-like algorithm.

We will follow the approach of Jacobi [24], in which the simultaneous diagonalization algorithm consists of a sequence of similarity transformations by plane rotations. A plane rotation in the (i, j) plane is a unitary matrix $R = R(i, j, c, s)$ of the form

$$R = R(i,j,c,s) = I + (c-1)e_i e_i^T - \bar{s} e_i e_j^T + s e_j e_i^T + (\bar{c}-1) e_j e_j^T, \tag{1}$$

where c, $s \in \mathbf{C}$ satisfy $|c|^2 + |s|^2 = 1$. For computational convenience we will often restrict c to be real and nonnegative.

A natural measure of the distance of the pair $(A, B) \in \mathbf{C}^{n\times n} \times \mathbf{C}^{n\times n}$ from diagonality is

$$\text{off}_2(A,B) = \sum_{i\neq j} |a_{ij}|^2 + \sum_{i\neq j} |b_{ij}|^2. \tag{2}$$

If p, q, r, $s \in \mathbf{C}$ and $ps - rq \neq 0$, then A and B are diagonal if and only if $(pA + qB)$ and $(rA + sB)$ are diagonal. Hence for each ordered quadruple $(p, q, r, s) \in \mathbf{C}^4$ such that $ps - rq \neq 0$, (2) gives rise to an alternative measure

$$\mathrm{off}_{2-pqrs} = \mathrm{off}_2(pA + qB, rA + sB). \tag{3}$$

Scaling A and B as suggested in [21] is the special case $q = r = 0$.

In broad outline, the algorithm we present consists of a sequence of similarity transformations by plane rotations, each of which is chosen to minimize either measure (2) or measure (3). This is done as follows. Let $R = R(i, j, c, s) \in \mathbf{C}^{n \times n}$ be a plane rotation with $c \in \mathbf{R}$. The restriction that c be real does not change the amount by which $\mathrm{off}_2(A, B)$ can be reduced. A simple calculation shows that

$$\begin{aligned}\mathrm{off}_2(RAR^H, RBR^H) = {} & \mathrm{off}_2(A, B) - |a_{ij}|^2 - |b_{ij}|^2 - |a_{ji}|^2 - |b_{ji}|^2 \\ & + |sc(\bar{a}_{ii} - \bar{a}_{jj}) + c^2\bar{a}_{ij} - s^2\bar{a}_{ji}|^2 \\ & + |sc(a_{ii} - a_{jj}) - s^2 a_{ij} + c^2 a_{ji}|^2 \\ & + |sc(\bar{b}_{ii} - \bar{b}_{jj}) + c^2\bar{b}_{ij} - s^2\bar{b}_{ji}|^2 \\ & + |sc(b_{ii} - b_{jj}) - s^2 b_{ij} + c^2 b_{ji}|^2.\end{aligned} \tag{4}$$

The choice of c and s that minimizes $\mathrm{off}_2(RAR^H, RBR^H)$ is the choice that minimizes

$$f_{ij}(c, s) = \|M_{ij}z\|_2 =_{\mathrm{def}} \left\| \begin{bmatrix} \bar{a}_{ij} & \frac{\bar{a}_{ii} - \bar{a}_{jj}}{\sqrt{2}} & -\bar{a}_{ji} \\ a_{ji} & \frac{a_{ii} - a_{jj}}{\sqrt{2}} & -a_{ij} \\ \bar{b}_{ij} & \frac{\bar{b}_{ii} - \bar{b}_{jj}}{\sqrt{2}} & -\bar{b}_{ji} \\ b_{ji} & \frac{b_{ii} - b_{jj}}{\sqrt{2}} & -b_{ij} \end{bmatrix} \begin{bmatrix} c^2 \\ \sqrt{2}cs \\ s^2 \end{bmatrix} \right\|_2 . \tag{5}$$

This is a constrained minimization problem. The constraint $|c|^2 + |s|^2 = 1$ implies that $\|z\|_2 = 1$. So, for all $z \in \mathbf{R}^3$, $z^H z = 1$,

$$\|M_{ij}z\|_2 \geq \min_{\|x\|_2 = 1} \|M_{ij}x\| = \sigma_{\min}(M_{ij}).$$

By parameterizing c and s as $c = \cos(\theta)$, $s = e^{i\phi}\sin(\theta)$ for θ, $\phi \in \mathbf{R}$ makes the minimization of (5) into a two-real-variable optimization problem. The following lemma shows that only the values $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]$ and $\phi \in [-\pi, \pi]$ need to be searched to minimize (5).

LEMMA 2.1. *All values of* (5) *occur with* $c = \cos(\theta)$, $s = e^{i\phi}\sin(\theta)$ *for some value of* $(\theta, \phi) \in [-\frac{\pi}{4}, \frac{\pi}{4}] \times [-\pi, \pi]$.

Proof. Define $g_{ij}(\theta, \phi)$ by $g_{ij}(\theta, \phi) = f_{ij}(\cos(\theta), e^{i\phi}\sin(\theta))$. The functions $\cos^2(\theta)$, $\cos(\theta)\sin(\theta)$, and $\sin^2(\theta)$ are π-periodic. Thus (5) implies that $g_{ij}(\theta, \phi)$ is π-periodic in θ. The special structure of M_{ij} in (5), along with the trivial observation $\left\|e^{-2i\phi}M_{ij}z\right\| = \|M_{ij}z\|$, shows that $g(\theta, \phi) = g(\theta + \frac{\pi}{2}, \phi) = g(-\theta - \frac{\pi}{2}, \phi)$. Hence, for each fixed value of ϕ, $g_{ij}(\theta, \phi)$ assumes all its values on $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]$. □

In the terminology of [40], an "inner rotation" is one for which $|c| \geq |s|$. An "outer rotation" is one for which $|c| < |s|$. Corresponding to each inner rotation that minimizes (5), there is an outer rotation that also minimizes (5). By choosing $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]$, we have made the choice of using only inner rotations. Our proof of quadratic convergence depends on them.

We have found no simple explicit formulae for the minimizers of (5). However, explicit formulae are known in some special cases. Goldstein and Horwitz [21] and

Eberlein [13] give explicit formulae for the special case $A = A^H$ and $B = -B^H$. We give explicit formulae for other special cases in §6. General optimization algorithms like those described in [19] provide effective ways to find minimizers $(\theta, \phi) \in [-\frac{\pi}{4}, \frac{\pi}{4}] \times [-\pi, \pi]$. In our MATLAB [29] implementation we used the exact minimizer if an explicit formula for the exact minimum was available. Otherwise, we used a heuristic approximate minimizer described below.

The work needed to find a minimizer of (5) does not grow with n. The work required to perform a similarity transformation by a rotation is proportional to n. For large enough problems, the work of finding minimizers of (5) is negligible. However, general constrained optimization algorithms are relatively complicated. A computer implementation of Algorithm 1 (described below) would devote the majority of the code to the optimization problem. Fortunately, it is sufficient to approximate the minimizer. Consider the following strategy:

Let $c_1 \in \mathbf{R}$ and $s_1 \in \mathbf{C}$ be minimizers of

$$g_{ij}(c_1, s_1) = |s_1 c_1(\bar{a}_{ii} - \bar{a}_{jj}) + c_1^2 \bar{a}_{ij} - s_1^2 \bar{a}_{ji}|^2 + |s_1 c_1(a_{ii} - a_{jj}) - s_1^2 a_{ij} + c_1^2 a_{ji}|^2, \tag{6}$$

and let $c_2 \in \mathbf{R}$ and $s_2 \in \mathbf{C}$ be minimizers of

$$h_{ij}(c_2, s_2) = |s_2 c_2(\bar{b}_{ii} - \bar{b}_{jj}) + c_2^2 \bar{b}_{ij} - s_2^2 \bar{b}_{ji}|^2 + |s_2 c_2(b_{ii} - b_{jj}) - s_2^2 b_{ij} + c_2^2 b_{ji}|^2. \tag{7}$$

Explicit formulae for (c_1, s_1) and (c_2, s_2) appear in [13], [21]. Use as approximate minimizers the pair $(c, s) = (c_i, s_i)$, $i = 1, 2$, that yields the smaller value of (5). This strategy has worked well in practice, and the proof of local quadratic convergence presented in §3 goes through with little modification for this approximate minimizer.

The following algorithm summarizes the procedure for simultaneous diagonalization of a commuting pair of normal matrices.

```
ALGORITHM 1.
INPUT: $\epsilon > 0$; $A$, $B \in \mathbf{C}^{n \times n}$ such that $AB = BA$, $AA^H = A^H A$ and $BB^H = B^H B$
OUTPUT: $Q \in \mathbf{C}^{n \times n}$ such that $\mathrm{off}_2(Q^H AQ, Q^H BQ) \leq \epsilon(\|A\|_F + \|B\|_F)$ and $QQ^H = I$.
1. $Q \leftarrow I$
2. WHILE $\mathrm{off}_2(A, B) > \epsilon(\|A\|_F + \|B\|_F)$
    3. FOR $i = 1, 2, 3, \ldots, n$
        4. FOR $j = i + 1, i + 2, i + 3, \ldots, n$
            5. Select $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]$ and $\phi \in [-\pi, \pi]$ such that $c = \cos(\theta)$ and $s = e^{i\phi}\sin(\theta)$ minimizes (5). (Or approximately minimizes (5) as described in (6) and (7).)
            6. $R \leftarrow R(i, j, c, s)$
            7. $Q \leftarrow QR$; $A \leftarrow R^H AR$; $B \leftarrow R^H BR$
```

If rotations are stored and applied in an efficient manner similar to that outlined for real rotations in [22, §5.1] or [33, §6.4], then each sweep (step 2) uses approximately $8n^3$ complex flops to update A and B and approximately $2n^3$ complex flops to accumulate Q. A complex flop is the computation effort required to execute the FORTRAN statement

(8)
```
A(I,J) = A(I,J) + S * A(K,J),
```

where `A` and `S` are of type `COMPLEX`.

Algorithm 1 needs storage for approximately $3n^2$ complex numbers. This can be shaved to $2n^2$ complex numbers if Q is not required.

As with the serial Jacobi algorithm, to promote rapid convergence in the case of multiple eigenvalues, it is a good idea to use a similarity transformation by a permutation matrix to put the diagonal entries of A and B in lexicographic order [20]. Such an eigenvalue ordering is required by our proof of local quadratic convergence in §3.

In our experience, for randomly chosen examples with $n \leq 80$, rarely does $\epsilon = 10^{-14}$ make Algorithm 1 require more than six sweeps. In §3 we show that Algorithm 1 has local quadratic convergence. There are, however, examples for which Algorithm 1 does not converge. For example, if $A \in \mathbf{R}^{n\times n}$ is given by

$$a_{kj} = \begin{cases} \cos((j+k)\pi/n) & \text{if } j \neq k, \\ \frac{2-n}{2}\cos(2k\pi/n) & \text{if } j = k, \end{cases} \tag{9}$$

and $B \in \mathbf{R}^{n\times n}$ is given by

$$b_{kj} = \begin{cases} \sin((j+k)\pi/n) & \text{if } j \neq k, \\ \frac{2-n}{2}\sin(2k\pi/n) & \text{if } j = k, \end{cases} \tag{10}$$

and $n > 6$, then it is easy to verify through Theorem 6.1 below that no rotation reduces $\operatorname{off}_2(A, B)$, so that Algorithm 1 with exact minimization in step 5 leaves A and B invariant. This example is essentially due to Voevodin [45].

In practice, rounding errors usually perturb (9) and (10) sufficiently to allow Algorithm 1 (with exact minimization in Step 5) to reduce $\operatorname{off}_2(A, B)$ and ultimately to converge to a simultaneously diagonal form, albeit rather slowly. For example, an early version of our experimental implementation on a Sun 4/60 computer (with unit round of approximately 10^{-16}) applied to the $n = 20$ case needed 47 sweeps to reduce $\operatorname{off}_2(A, B)$ to approximately $10^{-14}(\|A\| + \|B\|)$. Rounding errors did not perturb the $n = 30$ enough to allow convergence.

The example given by (9) and (10) is not held fixed by the heuristic minimization strategy, but it does sometimes require a few more sweeps than does a random example. Algorithm 1 with the approximate minimization heuristic in step 5 needed from 9 to 10 sweeps to reduce $\operatorname{off}_2(A, B)$ to less than $10^{-14}(\|A\| + \|B\|)$ for $n = 10$, $n = 15$, $n = 20$, $n = 25$, and $n = 30$.

Scaling one of the two commuting matrices by (say) replacing A by $A/2$, as suggested in [21], is a more reliable way to break away from a fixed point like that of (9) and (10). This is equivalent to using $\operatorname{off}_{2-1,0,0,2}$ in place of off_2. Motivated in part by the exceptional-shift strategy often used with the QR algorithm, we changed our experimental code to use $\operatorname{off}_{2-1,0,0,2}$ for one sweep, whenever $\operatorname{off}_2(A, B)$ declines by less than the 1% across a sweep. With this modification, only 6 to 7 sweeps were required to reduce $\operatorname{off}_2(A, B)$ to less than $10^{-14}(\|A\| + \|B\|)$ for the $n = 10$, $n = 15$, $n = 20$, $n = 25$, and $n = 30$ cases of (9) and (10). (The choice of 1% is ad hoc. A more cautious approach would require a greater per sweep reduction in $\operatorname{off}_2(A, B)$ and would try other scaling factors if $1/2$ does not work.)

We have not been able to show that Algorithm 1 with the above modification is globally convergent.

3. Convergence properties. Algorithm 1 shares many of the desirable properties of algorithms related to the serial Jacobi algorithm for the real symmetric eigenvalue problem [20], [23], [36], [38], [41], [46]. In our experience Algorithm 1, with the above strategy for avoiding stagnation, converges globally and ultimately quadratically.

In this section we establish the local quadratic convergence and numerical stability of Algorithm 1. In parenthetical remarks we give a rough sketch of how the argument can be modified to show the local convergence of the heuristic variation (6), (7).

Unfortunately, we do not have a proof of global convergence. Global convergence for special cases is shown in [21] and, by using a modified norm-reducing algorithm, in [17]. The latter, however, uses nonunitary transformations. In our case the use of unitary similarity transformations confers satisfactory numerical stability.

3.1. Local quadratic convergence. Here we follow [20], [36], [46] to show that Algorithm 1 ultimately converges quadratically. In particular, the resemblance to [36] is unmistakable.

Let the set of eigenvalues of A be $\{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n\}$, and let the set of eigenvalues of B be $\{\beta_1, \beta_2, \ldots, \beta_n\}$. Set

$$\delta = \frac{1}{2}\min\left(\min_{\alpha_i \neq \alpha_j} |\alpha_i - \alpha_j|, \min_{\beta_i \neq \beta_j} |\beta_i - \beta_j|\right). \tag{11}$$

Denote the kth rotation of Algorithm 1 by $R^{(k)} = R(i, j, c^{(k)}, s^{(k)})$, and let $A^{(k)}$ and $B^{(k)}$ be the values of A and B immediately before the kth rotation. Following common practice, we call $(a_{ij}^{(k)}, a_{ji}^{(k)}, b_{ij}^{(k)}, b_{ji}^{(k)})$ the "kth pivots," because the rotation $R^{(k)}(i, j, c^{(k)}, s^{(k)})$ is chosen to minimize

$$|a_{ij}^{(k+1)}|^2 + |a_{ji}^{(k+1)}|^2 + |b_{ij}^{(k+1)}|^2 + |b_{ji}^{(k+1)}|^2.$$

Define ρ_k by

$$\rho_k = \sqrt{\mathrm{off}_2(A^{(k)}, B^{(k)})}.$$

Partition $A^{(k)}$ and $B^{(k)}$ as

$$A^{(k)} = D_A^{(k)} + E_A^{(k)}$$

and

$$B^{(k)} = D_B^{(k)} + E_B^{(k)},$$

where $D_A^{(k)}$, $D_B^{(k)} \in \mathbf{C}^{n\times n}$ are diagonal and $E_A^{(k)}$, $E_B^{(k)} \in \mathbf{C}^{n\times n}$ have zero diagonals. Note that

$$\rho_k = \sqrt{\mathrm{off}_2(A^{(k)}, B^{(k)})} = \left\|[E_A^{(k)}, E_B^{(k)}]\right\|_F.$$

Suppose Algorithm 1 has converged to the point that

$$\rho_k = \sqrt{\mathrm{off}_2(A^{(k)}, B^{(k)})} < \frac{1}{2}\delta. \tag{12}$$

Using a permutation similarity, we may order the eigenvalues α_i and β_i so that

$$|a_{ii}^{(k)} - \alpha_i| \leq \left\|E_A^{(k)}\right\|_F \leq \rho_k < \frac{1}{2}\delta \tag{13}$$

and

$$|b_{ii}^{(k)} - \beta_i| \leq \left\|E_B^{(k)}\right\|_F \leq \rho_k < \frac{1}{2}\delta. \tag{14}$$

Because Algorithm 1 uses inner rotations and $\text{off}_2(A^{(k)}, B^{(k)})$ is monotonically decreasing, (13) and (14) hold throughout the remainder of Algorithm 1 [20], [36], [46]. In particular, the order in which the eigenvalues will eventually appear along the diagonals of A and B is fixed.

(The heuristic (6) and (7) does not necessarily make $\text{off}_2(A^{(k)}, B^{(k)})$ decrease monotonically. However, it can be shown that $\text{off}_2(A^{(k)}, B^{(k)})$ does not increase by more that $O(\delta^2)$. Thus under a slightly stronger hypothesis, (13) and (14) continue to hold.)

Each diagonal entry $a_{ii}^{(k)}$ is said to be "affiliated" with α_i, and each diagonal entry $b_{ii}^{(k)}$ is "affiliated" with β_i. As Algorithm 1 drives $\text{off}_2(A^{(k)}, B^{(k)})$ to zero, $\lim_{k\to\infty} a_{ii}^{(k)} = \alpha_i$ and $\lim_{k\to\infty} b_{ii}^{(k)} = \beta_i$. The affiliation of a diagonal entry does not change during subsequent steps of the algorithm [20], [36], [46].

Inequality (13) implies that if $\alpha_i \neq \alpha_j$, then

$$|a_{ii}^{(k)} - a_{jj}^{(k)}| > \delta. \tag{15}$$

Inequality (14) implies that if $\beta_i \neq \beta_j$, then

$$|b_{ii}^{(k)} - b_{jj}^{(k)}| > \delta. \tag{16}$$

An eigenvalue pair (α, β) has multiplicity m if there are m distinct integers i_j, $j = 1, 2, 3, \ldots, m$, for which $(\alpha, \beta) = (\alpha_{i_j}, \beta_{i_j})$. Using permutation similarity, we may order the diagonal entries so that affiliates of an m-fold multiple eigenvalue pair appear in m adjacent diagonal entries.

We will show the following theorem.

THEOREM 3.1. *Suppose that the hth rotation starts a sweep of Algorithm* 1, *i.e.,* $a_{12}^{(h)}$, $a_{21}^{(h)}$, $b_{12}^{(h)}$, $b_{21}^{(h)}$ *are the hth pivots. If* (12) *holds and if affiliates of m-fold multiple eigenvalue pairs appear in adjacent diagonal entries, then at the end of the sweep*

$$\rho_{h+n(n-1)/2} \leq 2n(9n - 13)\frac{\rho_h^2}{\delta}.$$

We will need several lemmas to prove Theorem 3.1. The proof has two parts. The first part establishes that

$$\sqrt{|a_{ij}^{(k+1)}|^2 + |a_{ji}^{(k+1)}|^2 + |b_{ij}^{(k+1)}|^2 + |b_{ji}^{(k+1)}|^2} \leq O(\rho_k^2). \tag{17}$$

In the second part we will show that subsequent rotations preserve (17) for subsequent values of k.

Lemmas similar to the following are well-known tools in the study of quadratic convergence of cyclic Jacobi algorithms [36], [41], [48].

LEMMA 3.2. *If* (12) *holds, if* $(\alpha_i, \beta_i) = (\alpha_j, \beta_j)$, *and if* $i \neq j$, *then*

$$\sqrt{|a_{ij}^{(k+1)}|^2 + |a_{ji}^{(k+1)}|^2 + |b_{ij}^{(k+1)}|^2 + |b_{ji}^{(k+1)}|^2} \leq \frac{\sqrt{2}\rho_k^2}{2\delta}.$$

Proof. Regardless of the particular choice of $R^{(k)}(i,j,c^{(k)},s^{(k)})$, the partitioning lemma [36, §2] implies that

$$\sqrt{|a_{ij}^{(k+1)}|^2 + |a_{ji}^{(k+1)}|^2} \leq \frac{\rho_k^2}{2\delta} \tag{18}$$

and

$$\sqrt{|b_{ij}^{(k+1)}|^2 + |b_{ji}^{(k+1)}|^2} \leq \frac{\rho_k^2}{2\delta}. \qquad \square \tag{19}$$

The next lemma covers the case $(\alpha_i, \beta_i) \neq (\alpha_j, \beta_j)$.

LEMMA 3.3. *Suppose that $a_{ij}^{(k)}$, $a_{ji}^{(k)}$, $b_{ij}^{(k)}$, $b_{ji}^{(k)}$, are the kth pivots. If* (12) *holds and if $(\alpha_i, \beta_i) \neq (\alpha_j, \beta_j)$, then*

$$\sqrt{|a_{ij}^{(k+1)}|^2 + |a_{ji}^{(k+1)}|^2 + |b_{ij}^{(k+1)}|^2 + |b_{ji}^{(k+1)}|^2} \leq 5\frac{\rho_k^2}{\delta}.$$

Proof. Without loss of generality assume that

$$|\alpha_i - \alpha_j| \geq |\beta_i - \beta_j|. \tag{20}$$

In particular, (20) implies $\alpha_i \neq \alpha_j$. In [36, §3] it is shown that there is a rotation $\hat{R} = R(i,j,c_A,s_A)$ such that for $\hat{A} = \hat{R}^H A^{(k)} \hat{R}$

$$\sqrt{|\hat{a}_{ij}|^2 + |\hat{a}_{ij}|^2} \leq \frac{\rho_k^2}{2\delta}. \tag{21}$$

In fact, this rotation minimizes the left-hand side of (21) (and (6)), so that

$$\|E_{\hat{A}}\|_F \leq \left\| E_A^{(k+1)} \right\|_F. \tag{22}$$

Set $\hat{B} = \hat{R}^H B^{(k)} \hat{R}$. Recall that $\hat{A}$ and $\hat{B}$ commute. The (i,j)th entry of the equation $\hat{A}\hat{B} - \hat{B}\hat{A} = 0$ is

$$\hat{b}_{ij}(\hat{a}_{ii} - \hat{a}_{jj}) - \hat{a}_{ij}(\hat{b}_{ii} - \hat{b}_{jj}) + \sum_{\substack{k \neq i \\ k \neq j}} \left(\hat{a}_{ik}\hat{b}_{kj} - \hat{b}_{ik}\hat{a}_{kj} \right) = 0. \tag{23}$$

Applying Hölder's inequality [35] to the sum and using (21) to bound the second term on the left-hand side of (23) gives

$$|\hat{b}_{ij}(\hat{a}_{ii} - \hat{a}_{jj})| \leq \rho_k^2 + \frac{\rho_k^2}{2\delta} |\hat{b}_{ii} - \hat{b}_{jj}|. \tag{24}$$

Inequality (22) implies that (13) and (15) hold for $\hat{A}$ and

$$\begin{aligned} |\hat{b}_{ii} - \hat{b}_{jj}| &\leq |b_{ii}^{(k)} - b_{jj}^{(k)}| + 2|b_{ij}^{(k)}| + 2|b_{ji}^{(k)}| \\ &\leq |b_{ii}^{(k)} - b_{jj}^{(k)}| + 4\rho_k \\ &\leq |b_{ii}^{(k)} - b_{jj}^{(k)}| + 2\delta. \end{aligned} \tag{25}$$

Inequalities (15) and (24) imply

$$|\hat{b}_{ij}| \leq \frac{\rho_k^2}{2\delta} \left(2 + \frac{\hat{b}_{ii} - \hat{b}_{jj}}{\hat{a}_{ii} - \hat{a}_{jj}} \right).$$

To bound the right-hand side, apply (25) to get

$$\left|\frac{\hat{b}_{ii}-\hat{b}_{jj}}{\hat{a}_{ii}-\hat{a}_{jj}}\right| \le \frac{|b_{ii}^{(k)}-b_{jj}^{(k)}|+2\delta}{|\hat{a}_{ii}-\hat{a}_{jj}|}.$$

The triangle inequality applied to $|(b_{ii}^{(k)}-\beta_i)+(\beta_i-\beta_j)+(\beta_j-b_{jj}^{(k)})|$ and $|(\alpha_i-\alpha_j)+(\alpha_i-\hat{a}_{ii})+(\alpha_j-\hat{a}_{jj})|$ gives

$$\left|\frac{\hat{b}_{ii}-\hat{b}_{jj}}{\hat{a}_{ii}-\hat{a}_{jj}}\right| \le \frac{|b_{ii}^{(k)}-\beta_i|+|\beta_i-\beta_j|+|\beta_j-b_{jj}^{(k)}|+2\delta}{|\alpha_i-\alpha_j|-|\alpha_i-\hat{a}_{ii}|-|\alpha_j-\hat{a}_{jj}|}.$$

Inequalities (13) and (14) imply

$$\left|\frac{\hat{b}_{ii}-\hat{b}_{jj}}{\hat{a}_{ii}-\hat{a}_{jj}}\right| \le \frac{|\beta_i-\beta_j|+3\delta}{|\alpha_i-\alpha_j|-\delta}.$$

Multiply and divide the right-hand side by $(\alpha_i-\alpha_j)^{-1}$, and apply (20), (13), and (14) to get

$$\left|\frac{\hat{b}_{ii}-\hat{b}_{jj}}{\hat{a}_{ii}-\hat{a}_{jj}}\right| \le \frac{\frac{\beta_i-\beta_j}{\alpha_i-\alpha_j}+\frac{3\delta}{\alpha_i-\alpha_j}}{1-\frac{\delta}{\alpha_i-\alpha_j}} \le \frac{1+\frac{3}{2}}{1-\frac{1}{2}} \le 5.$$

It now follows that $|\hat{b}_{ij}| \le \frac{7\rho_k^2}{2\delta}$. Similarly, it can be shown that $|\hat{b}_{ji}| \le \frac{7\rho_k^2}{2\delta}$. Therefore,

$$\sqrt{|\hat{a}_{ij}|^2+|\hat{a}_{ji}|^2+|\hat{b}_{ij}|^2+|\hat{b}_{ji}|^2} \le \frac{\sqrt{99}\rho_k^2}{2\delta} \le 5\frac{\rho_k^2}{\delta}. \tag{26}$$

Now, $R^{(k)}$ is chosen to minimize $|a_{ij}^{(k+1)}|^2+|a_{ji}^{(k+1)}|^2+|b_{ij}^{(k+1)}|^2+|b_{ji}^{(k+1)}|^2$, so that inequality (26) gives an upper bound on the minimum value. (If $R^{(k)}$ is chosen by using the heuristic (6), (7), then (26) is one of the two choices of $\operatorname{off}_2(A^{(k+1)},B^{(k+1)})$ over which the heuristic minimizes. Hence inequality (26) also gives an upper bound on the minimum value for this case.) □

Lemmas 3.2 and 3.3 imply (17). It remains to show that subsequent rotations preserve (17). The following lemma gives a bound on the angles of rotation that occur in Algorithm 1.

LEMMA 3.4. *If* (12) *holds and if* $(\alpha_i,\beta_i) \neq (\alpha_j,\beta_j)$, *then*

$$|s^{(k)}| \le \sqrt{2}\left(1+\frac{5\rho_k}{\delta}\right)\frac{\rho_k}{\delta} \le \frac{9\rho_k}{\delta}. \tag{27}$$

Proof. For ease of notation, set $(c,s)=(c^{(k)},s^{(k)})$. Without loss of generality we may assume that $\alpha_i \neq \alpha_j$. Lemma 3.3 implies that

$$|a_{ij}^{(k+1)}| = |sc(a_{ii}^{(k)}-a_{jj}^{(k)})-s^2a_{ji}^{(k)}+c^2a_{ij}^{(k)}| \le \frac{5\rho_k^2}{\delta}.$$

The triangle inequality gives

$$
\begin{aligned}
(28) \qquad |sc(a_{ii}^{(k)} - a_{jj}^{(k)})| &\le \frac{5\rho_k^2}{\delta} + |s^2 a_{ji}^{(k)} + c^2 a_{ij}^{(k)}| \\
&\le \frac{5\rho_k^2}{\delta} + |s^2 a_{ji}^{(k)}| + |c^2 a_{ij}^{(k)}| \\
&\le \frac{5\rho_k^2}{\delta} + \rho_k.
\end{aligned}
$$

Algorithm 1 chooses $c = \cos(\theta)$ for some $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]$, so that $|c| \ge \frac{\sqrt{2}}{2}$. Thus (15) and (28) together imply the left-hand inequality of (27). The right-hand inequality follows from the left-hand inequality and (12). □

Proof of Theorem 3.1. Fix a particular choice of (i, j), and consider $a_{ij}^{(k)}$ during a sweep of Algorithm 1 as k varies from $k = h$ through $k = h + n(n-1)/2$.

Note that ρ_k is monotonically decreasing in k, so if $(\alpha_i, \beta_i) = (\alpha_j, \beta_j)$, then Lemma 3.2 implies that for $k = h, h+1, h+2, \ldots, h + n(n-1)/2$,

$$|a_{ij}^{(k)}| \le \frac{\sqrt{2}\rho_k^2}{2\delta} \le \frac{\sqrt{2}\rho_h^2}{2\delta}.$$

(As noted in the proof of Lemma 3.2, a slightly weaker inequality holds for the heuristic (6), (7).)

Suppose that $(\alpha_i, \beta_i) \ne (\alpha_j, \beta_j)$ and that the kth pivots are $a_{ij}^{(k)}$, $a_{ji}^{(k)}$, $b_{ij}^{(k)}$, and $b_{ji}^{(k)}$. Fix i and j. Lemma 3.3 implies that $|a_{ij}^{(k+1)}| \le 5\rho_k^2/\delta$. Affiliates of multiple eigenvalue pairs appear in adjacent diagonal entries, so subsequent pivots in row i satisfy the hypothesis of Lemma 3.4. If $j < n$ and (for ease of notation) $(c, s) = (c^{(k+1)}, s^{(k+1)})$, then

$$
\begin{aligned}
|a_{ij}^{(k+2)}| &= |c a_{ij}^{(k+1)} + \bar{s} a_{j+1,j}^{(k+1)}| \\
&\le |a_{ij}^{(k+1)}| + |\bar{s}||a_{j+1,j}^{(k+1)}| \\
&\le 5\frac{\rho_k^2}{\delta} + \frac{9\rho_{k+1}^2}{\delta} \\
&\le 14\frac{\rho_k^2}{\delta} \le 14\frac{\rho_h^2}{\delta}.
\end{aligned}
$$

Similarly, each subsequent pivot in row i may increase $|a_{ij}^{(k)}|$ by at most $9\rho_h^2/\delta$.

Suppose that the last pivot in row i is the qth pivot. There are at most $n - i - 1 \le n - 2$ pivots in row i subsequent to the kth pivot. Hence for $j \ne i$, $|a_{ij}^{(q)}| \le (5 + 9(n-2))\,\rho_h^2/\delta$. We may bound the qth off-diagonal row sum by

$$r_{iq} =_{\text{def}} \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}^{(q)}|^2 \le (n-1)(5 + 9(n-2))^2 \frac{\rho_h^4}{\delta^2}.$$

Pivoting in rows other than row i leaves the off-diagonal row sum invariant, so that for $q < k < h + n(n-1)/2$, $r_{ik} = r_{iq} \le (n-1)(9n-13)^2 \rho_h^4/\delta^2$.

The same argument applied to B yields the identical bound for the off-diagonal row sums of B. Adding the bounds of the off-diagonal row sums of both A and B at the end of the sweep, we get

$$\rho_{h+n(n-1)/2} \le \sqrt{2n(n-1)}(9n-13)\frac{\rho_h^2}{\delta} \le 2n(9n-13)\frac{\rho_h^2}{\delta}. \qquad \square$$

It is clear from the derivation that the bound of Theorem 3.1 is not tight.

The theorem assumes that affiliates of multiple eigenvalue pairs appear in adjacent positions along the diagonal. This assumption guarantees that large angles of rotation do not cause fill-in of small off-diagonal entries. As suggested in [36], a threshold strategy would have much the same effect.

3.2. Rounding errors. Algorithm 1 uses only unitary similarity transformations. A classical rounding-error analysis of the construction and application of real rotations appears in [47, p. 131ff]. It is easily extended to the complex case. Applying it to Algorithm 1 results in the following theorem.

THEOREM 3.5. *Suppose the rotations in Algorithm* 1 *are constructed by using methods similar to those described in* [22, §5.1] *or* [33, §6.4]. *If* A, $B \in \mathbf{C}^{n\times n}$ *are a commuting normal pair input to Algorithm* 1 *and if* $Q^{(k)}$, $A^{(k)}$, $B^{(k)} \in \mathbf{C}^{n\times n}$ *are the computed versions of* Q, A, *and* B *after* k *sweeps, then there are matrices* F_A, $F_B \in \mathbf{C}^{n\times n}$ *such that*

$$(A+F_A)\,Q^{(k)} = Q^{(k)}A^{(k)},$$
$$(B+F_B)\,Q^{(k)} = Q^{(k)}B^{(k)},$$

and

$$\|F_A\|_F + \|F_B\|_F + \left\|{Q^{(k)}}^H Q - I\right\|_F \le p(n)\mu,$$

where μ *is the machine precision and* $p(n)$ *is a modest polynomial that depends on the details of the arithmetic.*

Thus the effects of rounding errors are equivalent to making small normwise perturbations in the original data matrices A and B. Algorithm 1 is backward numerically stable.

4. Real matrices. If Algorithm 1 is applied to a pair of real, commuting normal matrices A and B, then it will use nontrivial complex arithmetic on what is essentially a real problem. Thus storage must be set aside for two complex arrays instead of two real arrays, and complex arithmetic must be used instead of real arithmetic. Complex rounding errors may perturb the eigenvalues of A and B so that they do not appear in complex conjugate pairs. This section presents a modification of Algorithm 1 that avoids complex arithmetic.

Of course, there may be no real similarity transformation that simultaneously diagonalizes A and B. However, the following theorem shows that there is a real similarity transformation that block diagonalizes A and B with 1-by-1 and 2-by-2 blocks.

THEOREM 4.1. *If* A, $B \in \mathbf{R}^{n\times n}$, $A^TA = AA^T$, $B^TB = BB^T$, *and* $AB = BA$, *then there exists* $Q \in \mathbf{R}^{n\times n}$ *such that* $Q^TQ = I$, $D_A = Q^TAQ$ *is block diagonal and* $D_B = Q^TBQ$ *is block diagonal with* 2*-by-*2 *blocks and at most one trailing* 1*-by-*1 *block.*

Proof. If $n = 1$ or $n = 2$, then the theorem holds with $Q = [1] \in \mathbf{R}^{1\times 1}$ or $Q = I \in \mathbf{R}^{2\times 2}$.

Assume the induction hypothesis that the theorem holds for all real, normal, k-by-k, commuting pairs A and B for $k < n$ and $3 \le n$. Let $A \in \mathbf{R}^{n\times n}$ and $B \in \mathbf{R}^{n\times n}$ be commuting normal matrices, and let $x \in \mathbf{C}^n$ be a simultaneous eigenvector of A and B with eigenvalues λ_A and λ_B, respectively. There are two cases to consider: (i) x is linearly independent of $\bar{x}$ and (ii) x and $\bar{x}$ are linearly dependent.

If x and $\bar{x}$ are linearly independent, then $A\bar{x} = \bar{\lambda}_A\bar{x}$ and $B\bar{x} = \bar{\lambda}_B\bar{x}$. Thus x and $\bar{x}$ are simultaneous, linearly independent eigenvectors of A and B. The vectors $y = x+\bar{x}$

and $z = i(x - \bar{x})$ span a two-dimensional, real, invariant subspace of A and B. Let $U \in \mathbf{R}^{n\times n}$ be an orthogonal matrix whose first two columns form an orthonormal basis of $\operatorname{span}(y, z)$, and set $\tilde{A} = U^TAU$ and $\tilde{B} = U^TBU$. Thus $\tilde{A} = \operatorname{diag}(A_{11}, A_{22})$, $\tilde{B} = \operatorname{diag}(B_{11}, B_{22})$, where A_{11}, $B_{11} \in \mathbf{R}^{2\times 2}$ and A_{22}, $B_{22} \in \mathbf{R}^{n-2\times n-2}$. Observe that $\tilde{A}$ and $\tilde{B}$ are real, normal, commuting matrices, and hence A_{22} and B_{22} are commuting, normal $(n-2)$-by-$(n-2)$ matrices. By the induction hypothesis there exists an orthogonal matrix $V_{22} \in \mathbf{R}^{n-2\times n-2}$ such that $V_{22}^TA_{22}V_{22}$ and $V_{22}^TB_{22}V_{22}$ are block diagonal with 2-by-2 blocks and at most one trailing 1-by-1 block. If V_{11} is the 2-by-2 identity matrix and $V = \operatorname{diag}(V_{11}, V_{22})$, then the matrix $Q = UV$ satisfies the conclusion of the theorem.

If x and $\bar{x}$ are linearly dependent, then $\lambda_A \in \mathbf{R}$, $\lambda_B \in \mathbf{R}$ and x is a scalar multiple of some real, simultaneous eigenvector $y \in \mathbf{R}^n$. Without loss of generality we may choose y such that $\|y\|_2 = 1$. Let $U \in \mathbf{R}^{n\times n}$ be an orthogonal matrix whose last column is y, and set $\tilde{A} = U^TAU$ and $\tilde{B} = U^TBU$. Observe that $\tilde{A}$ and $\tilde{B}$ are real, normal, commuting matrices and that $\tilde{A} = \operatorname{diag}(A_{11}, a_{nn})$, $\tilde{B} = \operatorname{diag}(B_{11}, b_{nn})$, where $a_{nn} = \lambda_A$, $b_{nn} = \lambda_B$, and A_{11}, $B_{11} \in \mathbf{R}^{n-1\times n-1}$. By the induction hypothesis there exists an orthogonal matrix $V_{11} \in \mathbf{R}^{n-1\times n-1}$ such that $V_{11}^TA_{11}V_{11}$ and $V_{11}^TB_{11}V_{11}$ are block diagonal with 2-by-2 blocks and at most one trailing 1-by-1 block. If $V = \operatorname{diag}(V_{11}, [1])$, then the matrix $Q = UV$ satisfies the conclusion of the theorem. Note that at this point that if n is even, it is necessary to logically combine the two trailing 1-by-1 blocks into a single trailing 2-by-2 block. □

Partition each matrix $M \in \mathbf{R}^{n\times n}$ into a k-by-k block matrix as

$$(29) \qquad M = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1k} \\ M_{21} & M_{22} & \cdots & M_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ M_{k1} & M_{k2} & \cdots & M_{kk} \end{bmatrix},$$

where $k = \lfloor \frac{n+1}{2} \rfloor$ and for i, $j \le \lfloor \frac{n}{2} \rfloor$ $M_{ij} \in \mathbf{R}^{2\times 2}$. If n is odd, then for $j \le k-1$, $M_{kj} \in \mathbf{R}^{1\times 2}$, and for $i \le k-1$, $M_{ik} \in \mathbf{R}^{2\times 1}$ and $M_{kk} \in \mathbf{R}^{1\times 1}$. Theorem 4.1 states that there is a real orthogonal similarity transformation that simultaneously block diagonalizes A and B conformally with (29).

The next algorithm makes extensive use of elementary orthogonal matrices $Z = Z(i, j, U) \in \mathbf{R}^{n\times n}$ that are partitioned conformally with (29). Define the block rotation $Z(i, j, U)$ to have the form

$$(30) \qquad Z(i,j,U) = \begin{bmatrix} I &&&&&&&&&& \\ & \ddots &&&&&&&&& \\ && I &&&&&&&& \\ &&& Z_{ii} &&&& Z_{ij} &&& \\ &&&& I &&&&&& \\ &&&&& \ddots &&&&& \\ &&&&&& I &&&& \\ &&& Z_{ji} &&&& Z_{jj} &&& \\ &&&&&&&& I && \\ &&&&&&&&& \ddots & \\ &&&&&&&&&& I \end{bmatrix}.$$

Here,

$$\begin{bmatrix} Z_{ii} & Z_{ij} \\ Z_{ji} & Z_{jj} \end{bmatrix} =: U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}$$

is an orthogonal matrix. If $j < k$ or if n is even, then $U \in \mathbf{R}^{4\times 4}$. Otherwise, $U \in \mathbf{R}^{3\times 3}$. A similar definition of "blockwise rotation" appears in [18].

The Jacobi annihilators of [31] and the orthogonal matrices proposed in [42] are block rotations.

We will use block rotations $Z(i,j,U)$, where U has the form

$$\begin{aligned} U &= R(2,4,c_4,s_4)R(1,4,c_3,s_3)R(2,3,c_2,s_2)R(1,3,c_1,s_1) \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_4 & 0 & -s_4 \\ 0 & 0 & 1 & 0 \\ 0 & s_4 & 0 & c_4 \end{bmatrix} \begin{bmatrix} c_3 & 0 & 0 & -s_3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ s_3 & 0 & 0 & c_3 \end{bmatrix} \\ &\times \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_2 & -s_2 & 0 \\ 0 & s_2 & c_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_1 & 0 & -s_1 & 0 \\ 0 & 1 & 0 & 0 \\ s_1 & 0 & c_1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \end{aligned} \tag{31}$$

or

$$U = R(2,3,c_2,s_2)R(1,3,c_1,s_1) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c_2 & -s_2 \\ 0 & s_2 & c_2 \end{bmatrix} \begin{bmatrix} c_1 & 0 & -s_1 \\ 0 & 1 & 0 \\ s_1 & 0 & c_1 \end{bmatrix}, \tag{32}$$

where for $l = 1, 2, 3, 4$, c_l, $s_l \in \mathbf{R}$ and $c_l^2 + s_l^2 = 1$. It is convenient to parameterize c_l and s_l in (31) and (32) as $c_l = \cos(\theta_l)$ and $s_l = \sin(\theta_l)$, where $\theta_l \in \mathbf{R}$. Note that the Jacobi annihilators of [31] use a a different order of the factors in (31) and (32).

In broad outline, Algorithm 2 (presented below) is a block version of Algorithm 1 that uses the partition (29). It measures progress with a block version of (2) defined by

$$\mathrm{off}_B(A,B) = \sum_{i\neq j} \|A_{ij}\|_F^2 + \|B_{ij}\|_F^2,$$

where A and B are partitioned as in (29). At the (i,j)th step of a sweep it chooses a block rotation $Z = Z(i,j,U)$ to minimize $\mathrm{off}_B(Z^TAZ, Z^TBZ)$. It is easy to verify that for $i < j$

(33)

$$\begin{aligned} \mathrm{off}_B(Z^TAZ, Z^TBZ) = \mathrm{off}_B(A,B) &- \|A_{ij}\|_F^2 - \|A_{ji}\|_F^2 - \|B_{ij}\|_F^2 - \|B_{ji}\|_F^2 \\ &+ \left\|U_{12}^TA_{ii}U_{11} + U_{22}^TA_{ji}U_{11} + U_{12}^TA_{ij}U_{21} + U_{22}^TA_{jj}U_{21}\right\|_F^2 \\ &+ \left\|U_{11}^TA_{ii}U_{12} + U_{21}^TA_{ji}U_{12} + U_{11}^TA_{ij}U_{22} + U_{21}^TA_{jj}U_{22}\right\|_F^2 \\ &+ \left\|U_{12}^TB_{ii}U_{11} + U_{22}^TB_{ji}U_{11} + U_{12}^TB_{ij}U_{21} + U_{22}^TB_{jj}U_{21}\right\|_F^2 \\ &+ \left\|U_{11}^TB_{ii}U_{12} + U_{21}^TB_{ji}U_{12} + U_{11}^TB_{ij}U_{22} + U_{21}^TB_{jj}U_{22}\right\|_F^2. \end{aligned}$$

Minimizing (33) is equivalent to minimizing the sum of the last four terms. General optimization algorithms like those described in [19] provide effective ways to find minimizers $(c_l, s_l) = (\cos(\theta_l), \sin(\theta_l))$ for the parameters in (31) or (32).

For $U \in \mathbf{R}^{4\times 4}$ as above define $\tilde{U}$ by

$$\tilde{U} = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}.$$

Observe that $Z = Z(i,j,U)$ and $S = Z(i,j,\tilde{U})$ are both block rotations and that $\mathrm{off}_B(Z^TAZ, Z^TBZ) = \mathrm{off}_B(S^TAS, S^TBS)$. Thus we may choose a block rotation so that $\|U_{11}\|_F \geq \|U_{12}\|_F$. Such block rotations are sometimes called "inner" block rotations. The use of inner block rotations promotes convergence [18]. The 3-by-3 case is more complicated. However, the 3-by-3 case arises only if n is odd. As in [44], it can be eliminated simply by adding an extra row and column of zeros to A and B.

The following theorem implies that (33) is minimized by a block rotation of the form of (31) or (32). There is no need to use more general block rotations. A similar theorem about a different set of block rotations is proved in [42] by using different methods.

THEOREM 4.2.

1. *If $Q \in \mathbf{R}^{4\times 4}$ is orthogonal, then there is a block rotation $Z \in \mathbf{R}^{4\times 4}$ of the form of* (31) *and an orthogonal block diagonal matrix $D \in \mathbf{R}^{4\times 4}$ with* 2*-by-*2 *blocks such that $ZQ = D$.*
2. *If $Q \in \mathbf{R}^{3\times 3}$ is orthogonal, then there is a block rotation $Z \in \mathbf{R}^{3\times 3}$ of the form of* (32) *and an orthogonal block diagonal matrix $D \in \mathbf{R}^{3\times 3}$ with* 2*-by-*2 *and* 1*-by-*1 *blocks such that $ZQ = D$.*

Proof. We will prove statement 1. The proof of statement 2 is similar.

Select $c_1 = \cos(\theta_1)$ and $s_1 = \sin(\theta_1)$ such that

$$c_1 \det \begin{bmatrix} q_{21} & q_{22} \\ q_{31} & q_{32} \end{bmatrix} + s_1 \det \begin{bmatrix} q_{21} & q_{22} \\ q_{11} & q_{12} \end{bmatrix} = 0.$$

Set $Q^{(1)} = R(1,3,c_1,s_1)Q$. With this choice of c_1 and s_1

$$\det \begin{bmatrix} q_{21}^{(1)} & q_{22}^{(1)} \\ q_{31}^{(1)} & q_{32}^{(1)} \end{bmatrix} = 0,$$

so that there is a choice of $c_2 = \cos(\theta_2)$ and $s_2 = \sin(\theta_2)$ such that

$$s_2 q_{21}^{(1)} + c_2 q_{31}^{(1)} = s_2 q_{22}^{(1)} + c_2 q_{32}^{(1)} = 0.$$

Thus $Q^{(2)} = R(2,3,c_2,s_2)Q^{(1)}$ has zeros in the $(3,1)$ and $(3,2)$ entries. Select $c_3 = \cos(\theta_3)$ and $s_3 = \sin(\theta_3)$ such that

$$c_3 \det \begin{bmatrix} q_{21}^{(2)} & q_{22}^{(2)} \\ q_{41}^{(2)} & q_{42}^{(2)} \end{bmatrix} + s_3 \det \begin{bmatrix} q_{21}^{(2)} & q_{22}^{(2)} \\ q_{11}^{(2)} & q_{12}^{(2)} \end{bmatrix} = 0.$$

Set $Q^{(3)} = R(1,4,c_3,s_3)Q^{(2)}$. With this choice of c_3 and s_4 we get

$$\det \begin{bmatrix} q_{21}^{(3)} & q_{22}^{(3)} \\ q_{41}^{(3)} & q_{42}^{(3)} \end{bmatrix} = 0.$$

Hence there is a choice of $c_4 = \cos(\theta_4)$ and $s_4 = \sin(\theta_4)$ such that

$$s_2 q_{21}^{(1)} + c_2 q_{41}^{(1)} = s_2 q_{22}^{(1)} + c_2 q_{42}^{(1)} = 0.$$

Thus $Q^{(4)} = R(2,4,c_4,s_4)Q^{(3)}$ has zeros in the $(3,1)$, $(3,2)$, $(4,1)$, and $(4,2)$ entries.

Set

$$Z = R(2,4,c_4,s_4)R(1,4,c_3,s_3)R(2,3,c_2,s_2)R(1,3,c_1,s_1),$$

and set $D = Q^{(4)} = ZQ$. Observe that D is an orthogonal, block upper-triangular matrix with 2-by-2 blocks. Hence D is block diagonal with 2-by-2 blocks. □

The following algorithm summarizes the procedure for simultaneous diagonalization of a commuting pair of real normal matrices.

```
ALGORITHM 2.
INPUT: $\epsilon > 0$; $A$, $B \in \mathbf{R}^{n\times n}$ such that $AB = BA$, $AA^T = A^TA$, and $BB^T = B^TB$
OUTPUT: $Q \in \mathbf{R}^{n\times n}$ such that $\mathrm{off}_B(Q^TAQ, Q^TBQ) \le \epsilon(\|A\|_F+\|B\|_F)$ and $QQ^T = I$.
1. $Q \leftarrow I$
2. WHILE $\mathrm{off}_B(A,B) > \epsilon(\|A\|_F + \|B\|_F)$
    3. FOR $i = 1, 2, 3, \ldots, \lfloor \frac{n+1}{2} \rfloor$
        4. FOR $j = i+1, i+2, i+3, \ldots, \lfloor \frac{n+1}{2} \rfloor$
            5. Select an inner block rotation $Z \leftarrow Z(i,j,U)$ that minimizes $\mathrm{off}_B(Z^TAZ, Z^TBZ)$
            6. $Q \leftarrow QZ$; $A \leftarrow Z^TAZ$; $B \leftarrow Z^TBZ$
```

If rotations are stored and applied in the efficient manner described in [22, §5.1] or [33, §6.4], then each sweep (step 2) uses approximately $8n^3$ real flops to update A and B and approximately $2n^3$ real flops to accumulate Q. A real flop is the computation effort required to execute FORTRAN statement (8) if `A` and `S` are of type `REAL`. The rough estimate that one complex flop is equivalent to four real flops implies that Algorithm 2 does about half the work of Algorithm 1.

Algorithm 1 also needs storage for approximately $3n^2$ real numbers. This can be shaved to $2n^2$ real numbers if Q is not required.

5. Quaternions. A matrix $H \in \mathbf{C}^{2n\times 2n}$ has *quaternion* structure if it is of the form

$$JH = \bar{H}J, \tag{34}$$

where $J \in \mathbf{R}^{2n\times 2n}$ is defined by

$$J = \mathrm{diag}(E, E, E, \ldots, E) \tag{35}$$

and

$$E = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

If a quaternion matrix H is partitioned into an n-by-n block matrix with 2-by-2 blocks, then each block is of the form

$$H_{ij} = \begin{bmatrix} u_{ij} & -\bar{v}_{ij} \\ v_{ij} & \bar{u}_{ij} \end{bmatrix}, \tag{36}$$

where u_{ij}, $v_{ij} \in \mathbf{C}$. It is easy to show that (34) and (36) are equivalent.

If $P \in \mathbf{R}^{2n\times 2n}$ is the permutation matrix

$$P = [e_1, e_3, e_5, \ldots, e_{2n-1}, e_2, e_4, e_6, \ldots, e_{2n}],$$

then

$$PHP^T = \begin{bmatrix} U & -\bar{V} \\ V & \bar{U} \end{bmatrix}. \tag{37}$$

In other treatments, e.g., [7], quaternion matrices are defined to have the form of (37), but in this context (36) makes the explication somewhat simpler. Quaternion matrices arise naturally from quantum mechanical problems that have time-reversal symmetry [12], [26], [34], and in eigenvalue problems that have two special structures [6].

The eigenvalues and eigenvectors of quaternion matrices appear in pairs. If (λ, x) is an eigenvalue–eigenvector pair of a quaternion matrix, then $(\bar{\lambda}, J\bar{x})$ is also an eigenvalue–eigenvector pair. A quaternion matrix H has a quaternion Schur decomposition as $HQ = QT$, where $Q \in \mathbf{C}^{2n\times 2n}$ is a unitary quaternion matrix and $T \in \mathbf{C}^{2n\times 2n}$ is an upper-triangular quaternion matrix [7]. If $H \in \mathbf{C}^{2n\times 2n}$ is also normal, then T is a diagonal quaternion matrix. From this it is easy to show the following theorem.

THEOREM 5.1. *If A, $B \in \mathbf{C}^{2n\times 2n}$ are normal commuting quaternion matrices, then there are a unitary quaternion matrix $Q \in \mathbf{C}^{2n\times 2n}$ and quaternion diagonal matrices $D_A, D_B \in \mathbf{C}^{2n\times 2n}$ such that $AQ = QD_A$ and $BQ = QD_B$.*

Algorithm 1 applied to a commuting pair of normal quaternion matrices will destroy the quaternion structure. General $2n$-by-$2n$ matrices must be stored and updated. Rounding errors may destroy the pairing of eigenvalues and eigenvectors. Algorithm 1 takes no advantage of the pairing of eigenvalues.

Theorem 5.1 suggests that if quaternion structure were preserved throughout Algorithm 1, then work and storage requirements would be cut in half and the special pairing of eigenvalues and eigenvectors will be preserved despite rounding errors.

The key to preserving quaternion structure is the observation that products and inverses of quaternion matrices are quaternion matrices. In particular, there is a rich class of quaternion unitary matrices to use in a modified version of Algorithm 1.

A quaternion rotation is a quaternion unitary matrix $W \in \mathbf{C}^{2n\times 2n}$ of the form $W = W_1W_2W_3$, where

$$W_1 = R(2i, 2j, \bar{c}_4, \bar{s}_4)R(2i-1, 2j-1, c_4, s_4), \tag{38}$$

$$W_2 = R(2j-1, 2j, c_3, s_3)R(2i-1, 2i, c_2, s_2), \tag{39}$$

and

$$W_3 = R(2i, 2j, \bar{c}_1, \bar{s}_1)R(2i-1, 2j-1, c_1, s_1). \tag{40}$$

The matrices W_1, W_2, W_3 are quaternion unitary matrices. (The two rotations in (39) are quaternion, but the individual rotations in (38) and (40) are not.) Hence W is quaternion and unitary. In terms of the block 2-by-2 partitioning (29), for

$1 \leq i < j \leq k = n$ the quaternion rotation $W = W(i, j, U)$ is of the form

$$W = \begin{bmatrix} I &&&&&&&&&& \\ & \ddots &&&&&&&&& \\ && I &&&&&&&& \\ &&& W_{ii} &&&& W_{ij} &&& \\ &&&& I &&&&&& \\ &&&&& \ddots &&&&& \\ &&&&&& I &&&& \\ &&& W_{ji} &&&& W_{jj} &&& \\ &&&&&&&& I && \\ &&&&&&&&& \ddots & \\ &&&&&&&&&& I \end{bmatrix}, \tag{41}$$

where

$$\begin{aligned} \begin{bmatrix} W_{ii} & W_{ij} \\ W_{ji} & W_{jj} \end{bmatrix} &= \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \\ &= R(2,4,\bar{c}_4,\bar{s}_4)R(1,3,c_4,s_4) \\ &\quad \times R(3,4,c_3,s_3)R(1,2,c_2,s_2) \\ &\quad \times R(2,4,\bar{c}_1,\bar{s}_1)R(1,3,c_1,s_1). \end{aligned} \tag{42}$$

Up to column scaling, quaternion unitary matrices are products of quaternion rotations. This was proved for symplectic matrices in the permuted form (37) in [32], but the methods carry over to the quaternion case with few modifications.

Suppose A, $B \in \mathbf{C}^{2n \times 2n}$ form a commuting pair of normal quaternion matrices. In broad outline, Algorithm 3 (presented below) is a block version of Algorithm 1 partitioned into 2-by-2 blocks. It measures progress with a block version of (2) defined by

$$\text{off}_Q(A, B) = \sum_{i \neq j} \|A_{ij}\|_F^2 + \|B_{ij}\|_F^2.$$

At the (i, j)th step of a sweep it uses a quaternion rotation $W = W(i, j, U)$ to minimize $\text{off}_Q(W^H A W, W^H B W)$.

By partitioning U as in (42), it is easy to verify that for $i < j$

$$\begin{aligned} \text{off}_Q(W^H AW, W^H BW) &= \text{off}_Q(A, B) - \|A_{ij}\|_F^2 - \|A_{ji}\|_F^2 - \|B_{ij}\|_F^2 - \|B_{ji}\|_F^2 \\ &\quad + \left\|U_{12}^H A_{ii} U_{11} + U_{22}^H A_{ji} U_{11} + U_{12}^H A_{ij} U_{21} + U_{22}^H A_{jj} U_{21}\right\|_F^2 \\ &\quad + \left\|U_{11}^H A_{ii} U_{12} + U_{21}^H A_{ji} U_{12} + U_{11}^H A_{ij} U_{22} + U_{21}^H A_{jj} U_{22}\right\|_F^2 \\ &\quad + \left\|U_{12}^H B_{ii} U_{11} + U_{22}^H B_{ji} U_{11} + U_{12}^H B_{ij} U_{21} + U_{22}^H B_{jj} U_{21}\right\|_F^2 \\ &\quad + \left\|U_{11}^H B_{ii} U_{12} + U_{21}^H B_{ji} U_{12} + U_{11}^H B_{ij} U_{22} + U_{21}^H B_{jj} U_{22}\right\|_F^2. \end{aligned} \tag{43}$$

Minimizing (43) is equivalent to minimizing the sum of the last four terms. General optimization algorithms like those described in [19] provide effective ways to find minimizers $(c_l, s_l) = (\cos(\theta_l), \sin(\theta_l))$ for the parameters in (38), (39), and (40).

For $U \in \mathbf{R}^{4\times 4}$ partitioned as in (42) define $\tilde{U}$ by

$$\tilde{U} = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}.$$

Observe that $W = W(i,j,U)$ and $S = W(i,j,\tilde{U})$ are both quaternion rotations and that $\text{off}_Q(W^HAW, W^HBW) = \text{off}_Q(S^HAS, S^HBS)$. Thus we may choose the quaternion rotation so that $\|U_{11}\|_F \geq \|U_{21}\|_F$. Such block rotations are sometimes called "inner" block rotations. The use of inner block rotations promotes convergence [18].

The following modification of Algorithm 1 summarizes the procedure for simultaneous diagonalization of a commuting pair of quaternion normal matrices. Note that the output of Algorithm 3 includes a pair of block diagonal matrices with 2-by-2 blocks. The trivial step of simultaneously diagonalizing the commuting 2-by-2 blocks is omitted.

```
ALGORITHM 3.
INPUT: $\epsilon > 0$; $A, B \in \mathbf{C}^{2n\times 2n}$ such that $AB = BA$, $JA = \bar{A}J$, $JB = \bar{B}J$,
    $AA^H = A^HA$ and $BB^H = B^HB$
OUTPUT: $Q \in \mathbf{C}^{2n\times 2n}$ such that $\text{off}_Q(Q^HAQ, Q^HBQ) \leq \epsilon(\|A\|_F + \|B\|_F)$, $JQ =$
    $\bar{Q}J$, and $QQ^H = I$
1. $Q \leftarrow I$
2. WHILE $\text{off}_Q(A,B) > \epsilon(\|A\|_F + \|B\|_F)$
    3. FOR $i = 1, 2, 3, \ldots, \frac{n}{2}$
        4. FOR $j = i+1, i+2, i+3, \ldots, \frac{n}{2}$
            5. Select an inner quaternion rotation $W \leftarrow W(i,j,U)$ that mini-
               mizes $\text{off}_Q(W^HAW, W^HBW)$
            6. $Q \leftarrow QZ$; $A \leftarrow W^HAW$; $B \leftarrow W^HBW$
```

Note that Q is a product of quaternion rotations, so that Q is a quaternion unitary matrix. The quaternion structure of A and B is preserved throughout. Equations (36) and (37) show that only $2n^2$ complex numbers are needed to represent a quaternion matrix. With this economy Algorithm 3 needs approximately $6n^2$ storage. If rotations are stored and applied in an efficient manner similar to [22, §5.1] or [33, §6.4], then each sweep (step 2) uses approximately $24n^3$ complex flops to update A and B and approximately $6n^3$ complex flops to accumulate Q.

6. Additional symmetry structure.

In [6] simultaneous diagonalization problems usually have an additional special structure in addition to normality. At times A and B are Hermitian, real symmetric, real skew symmetric, or quaternion. In this section we outline how to modify Algorithm 1 to take advantage of additional special structure.

6.1. Hermitian case.

If $(A,B) \in \mathbf{C}^{n\times n} \times \mathbf{C}^{n\times n}$ is a pair of commuting Hermitian matrices, then A is the Hermitian part and iB is the skew-Hermitian part of the normal matrix $H = A+iB$. All normal matrices are of this form. Algorithm 1 reduces to the Jacobi method for normal matrices for which explicit formulae for minimizers of (5) are known [21].

6.2. Real symmetric case.

The following theorem shows that if A and B are real symmetric, then no complex arithmetic is required by Algorithm 1. Moreover, there is an explicit formula for the minimizer of (5).

THEOREM 6.1. *Suppose A, $B \in \mathbf{R}^{n\times n}$, $A = A^T$, $B = B^T$, and $AB = BA$. If c, $s \in \mathbf{R}$, $c^2 + s^2 = 1$, and*

$$w = \begin{bmatrix} c^s - s^2 \\ 2cs \end{bmatrix}$$

is a right singular vector of

$$L_{ij} = \begin{bmatrix} a_{ij} & \dfrac{a_{ii} - a_{jj}}{2} \\ b_{ij} & \dfrac{a_{ii} - b_{jj}}{2} \end{bmatrix}$$

corresponding to its smallest singular value, then (c, s) is a minimizer of (5). Furthermore, if L_{ij} has distinct singular values and $(\tilde{c}, \tilde{s}) \in \mathbf{C} \times \mathbf{C}$ minimizes (5), then

$$\tilde{w} = \begin{bmatrix} \tilde{c}^s - \tilde{s}^2 \\ 2\tilde{c}s \end{bmatrix}$$

is a scalar multiple of w.

Proof. In this case, (5) simplifies to

$$f_{ij}(c, s) = \sqrt{2}\|L_{ij}w\|_2 =_{\text{def}} \sqrt{2}\left\| \begin{bmatrix} a_{ij} & \dfrac{a_{ii} - a_{jj}}{2} \\ b_{ij} & \dfrac{a_{ii} - b_{jj}}{2} \end{bmatrix} \begin{bmatrix} c^2 - s^2 \\ 2cs \end{bmatrix} \right\|_2 . \tag{44}$$

If the double-angle formulae for $s = \sin(\theta)$ and $c = \cos(\theta)$ are used, any two-dimensional real vector of Euclidean length 1 may be written in the form of w for some choice of c, $s \in \mathbf{R}$, $c^2 + s^2 = 1$, and $c \geq s$. In particular, there is a choice of c and s such that w is a right singular vector of L_{ij} that corresponds to the smallest singular value of L_{ij}. Note that L_{ij} is real, so that w may be chosen to be real. Clearly, (44) is bounded below by $\sigma_{\min}(L_{ij})$, so that this choice of c and s minimizes $f_{ij}(c, s)$. If L_{ij} has distinct singular values, then its singular vectors are unique up to scalar multiples. □

In addition to simplifying the calculation of the minimizing c and s, Theorem 6.1 shows that only real rotations are needed. The real symmetric structure of A and B is preserved throughout. Consequently, only the real part of the upper triangle of A and B and the real part of Q need to be stored and updated. These modifications cut storage requirements of Algorithm 1 by a factor of 3. The work requirements are cut to $6n^3$ real flops per sweep.

6.3. Real symmetric–real skew-symmetric case. If $A \in \mathbf{R}^{n\times n}$ is symmetric, $B \in \mathbf{R}^{n\times n}$ is skew symmetric, and $AB = BA$, then $H = A + B \in \mathbf{R}^{n\times n}$ is normal. All real normal matrices are of this form. In this case, Algorithm 2 applies with $A := A + B$ and $B := 0$. Note that $B := 0$ is invariant throughout Algorithm 2, so that no work need be expended for updating B and no storage need be allocated for storing B. This cuts the cost of Algorithm 2 down to approximately $4n^3$ real flops per sweep for updating $A := A + B$ and approximately $2n^3$ flops per sweep for updating Q. A similar algorithm that uses a different set of block rotations is suggested in [42].

6.4. Real skew-symmetric–real skew-symmetric case. The skew-symmetric structure of A and B is preserved throughout Algorithm 2. It is necessary to store and update only the upper-triangular part of A and B. This cuts the work required by Algorithm 2 to the same level as that required for the symmetric–skew-symmetric case.

6.5. Hermitian quaternion case. If (A, B) are Hermitian quaternion matrices, then A is the Hermitian part and iB is the skew-Hermitian part of the normal matrix $H = A + iB$. However, H is not quaternion. (The matrix iB is antiquaternion [7]. See §6.6 below.) Nevertheless, Algorithm 3 preserves both the quaternion and Hermitian structures, so that only the upper triangles of A and B need be stored and updated. Thus both the storage required for A and B and the work required for updating A and B are cut in half.

6.6. Hermitian quaternion–Hermitian antiquaternion case. A matrix $H \in \mathbf{C}^{2n\times 2n}$ is said to be antiquaternion if and only if iH is quaternion. If $A \in \mathbf{C}^{2n\times 2n}$ is Hermitian and quaternion and $B \in \mathbf{C}^{2n\times 2n}$ is Hermitian and antiquaternion, then A is the Hermitian part and iB is the skew-Hermitian part of the quaternion normal matrix $H = A + iB$. Moreover, all quaternion normal matrices are of this form. Algorithm 3 applies with $A := A + iB$ and $B := 0$. Note that $B := 0$ is invariant, so that it need not be stored or updated. This economy cuts the work and storage requirements of Algorithm 3 down to approximately $18n^3$ complex flops per sweep. If Q is not required, then the work requirement drops to $12n^3$ complex flops per sweep.

6.7. Commuting m-tuples. Algorithm 1 extends to the problem of simultaneously diagonalizing m commuting normal matrices. In this case the minimization problem (5) uses a $2m$-by-3 matrix coefficient matrix. Algorithm 1 must apply similarity transformations to all m commuting matrices.

7. Conclusions. We have presented a Jacobi-like algorithm for simultaneous diagonalization of commuting pairs of complex normal matrices. The algorithm uses a sequence of similarity transformations by elementary complex rotations to drive the off-diagonal entries to zero. Convergence and rounding-error properties are similar to those of the serial Jacobi algorithm [20], [38], [46]. We have shown that its asymptotic convergence rate is quadratic and that it is numerically stable in the sense that the computed eigenvalues and eigenvectors are correct for a perturbation of the data. Empirically, it appears to converge globally, but we have not been able to give a proof. The algorithm can easily be modified to preserve and exploit the additional special structure of real matrices, quaternion matrices, and real symmetric matrices.

Acknowledgments. We thank Gabriele Steinmeyer for help in testing some of the algorithms and Vjeran Hari for comments on an early draft of this paper. We also thank an anonymous referee for bringing [45] to our attention, and we thank K. Veselic for providing us with a copy of the paper and a translation.

REFERENCES

[1] J. Barlow and J. Demmel, *Computing Accurate Eigensystems of Scaled Diagonally Dominant Matrices*, Tech. Report 421, Courant Institute of Mathematical Sciences, Department of Computer Science, New York University, New York, NY, 1988.

[2] M. Berry and A. H. Sameh, *Multiprocessor Jacobi algorithms for dense symmetric eigenvalue and singular value decompositions*, in Proc. 1986 International Conference on Parallel Processing, Washington, D.C., K. Hwant, S. M. Jacobi, and E. E. Swartzlander, eds., IEEE Computer Society and Assoc. for Computer Machinery, 1986, pp. 433–440.

[3] ———, *A Parallel Algorithm for the Singular Value and Dense Symmetric Eigenvalue Problem*, Tech. Report, Center for Supercomputing Research and Development, University of Illinois at Urbana-Champaign, Urbana, IL, 1988.

[4] C. H. Bischof, *The two-sided block Jacobi method on a hypercube architecture*, in Hypercube Multiprocessors, M. T. Heath, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987, pp. 612–618.

[5] R. P. Brent and F. T. Luk, *The solution of singular-value and symmetric eigenvalue problems on multiprocessor arrays*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 69–84.

[6] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, *A chart of numerical methods for structured eigenvalue problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 419–453.

[7] ———, *A quaternion* QR *algorithm*, Numer. Math., 55 (1989), pp. 83–95.

[8] K. Chen and K. Irani, *A Jacobi algorithm and its implementation on parallel computers*, in Proc. 18th Annual Allerton Conf. Communication Control and Computing, Monticello, IL, 1980.

[9] R. O. Davies and J. J. Modi, *A direct method for computing eigenproblem solutions on a parallel computer*, Linear Algebra Appl., 77 (1986), pp. 61–74.

[10] J. W. Demmel and W. Kahan, *Computing small singular values of bidiagonal matrices with guaranteed high relative accuracy*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 873–912.

[11] J. W. Demmel and K. Veselić, *Jacobi's Method is More Accurate than* QR, Tech. Report 468, Courant Institute of Mathematical Sciences, Department of Computer Science, New York University, New York, NY, 1989.

[12] J. J. Dongarra, J. R. Gabriel, D. D. Kölling, and J. H. Wilkinson, *The eigenvalue problem for Hermitian matrices with time reversal symmetry*, Linear Algebra Appl., 60 (1984), pp. 27–42.

[13] P. J. Eberlein, *A Jacobi-like method for the automatic computation of eigenvalues and eigenvectors of an arbitrary matrix*, J. Soc. Indust. Appl. Math., 10 (1962), pp. 74–88.

[14] ———, *Solution to the complex eigenproblem by norm reducing Jacobi-type method*, Numer. Math., 14 (1970), pp. 232–245.

[15] ———, *On one-sided Jacobi methods for parallel computation*, SIAM J. Algebraic Discrete Meth., 8 (1987), pp. 790–796.

[16] ———, *On using the Jacobi method on the hypercube*, in Proc. 2nd Conference on Hypercube Multiprocessors, M. T. Heath, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987, pp. 605–611.

[17] A. Erdmann, *Über Jacobi-ähnliche Verfahren zur Lösung des Eigenwertproblems nichtnormaler komplexer Matrizen*, Dissertation, Fernuniversität Hagen, Hagen, Germany, 1984.

[18] K. V. Fernando and S. J. Hammarling, *On block Kogbetliantz methods for computation of the* SVD, in SVD and Signal Processing: Algorithms, Applications and Architectures, F. Deprettere, ed., North-Holland, Amsterdam, 1988.

[19] R. Fletcher, *Practical Methods of Optimization*, Vols. 1 and 2, John Wiley, New York, 1981.

[20] G. E. Forsythe and P. Henrici, *The cyclic Jacobi method for computing the principal values of a complex matrix*, Trans. Amer. Math. Soc., 94 (1960), pp. 1–23.

[21] H. H. Goldstein and L. P. Horwitz, *A procedure for the diagonalization of normal matrices*, J. Assoc. Comput. Mach., 6 (1959), pp. 176–195.

[22] G. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.

[23] V. Hari, *On Sharp Quadratic Convergence Bounds for the Serial Jacobi Methods*, Tech. Report, Department of Mathematics, University of Zagreb, Zagreb, Yugoslavia, 1989.

[24] C. G. Jacobi, *Über ein leichtes Verfahren, die in der Theorie der Säkulärstörungen vorkommendem Gleichungen numerisch aufzulösen*, J. Reine Angew. Math., 1846, pp. 51–95.

[25] W. Kahan, *Accurate Eigenvalues of a Symmetric Tridiagonal Matrix*, Tech. Report CS 41, Department of Computer Science, Stanford University, Stanford, CA, 1966; revised 1968.

[26] M. Lax, *Symmetry Principles in Solid State and Molecular Physics*, 1st ed., John Wiley and Sons, New York, 1974.

[27] F. T. Luk, *A triangular processor array for computing singular values*, Linear Algebra Appl., 77 (1986), pp. 259–273.

[28] F. T. Luk and H. Park, *A proof of convergence for two parallel Jacobi* SVD *algorithms*, IEEE Trans. Comput., 38 (1989), pp. 806–811.

[29] C. Moler, MATLAB *User's Guide*, Tech. Report CS81-1, Department of Computer Science, University of New Mexico, Albuquerque, NM, 1980.

[30] B. Noble and W. Daniel, *Applied Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[31] M. H. C. Paardekooper, *An eigenvalue algorithm for skew-symmetric matrices*, Numer. Math., 17 (1971), pp. 189–202.

[32] C. Paige and C. F. Van Loan, *A Schur decomposition for Hamiltonian matrices*, Linear Algebra Appl., 41 (1981), pp. 11–32.

[33] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[34] N. Rösch, *Time-reversal symmetry, Kramers' degeneracy and the algebraic eigenvalue problem*, Chem. Phys., 80 (1983), pp. 1–5.

[35] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[36] A. Ruhe, *On the quadratic convergence of the Jacobi method for normal matrices*, BIT, 7 (1967), pp. 305–313.

[37] A. H. Sameh, *On Jacobi and Jacobi–like algorithms for a parallel computer*, Math. Comp., 25 (1971), pp. 579–590.

[38] A. Schönhage, *Zur Konvergenz des Jacobi-Verfahrens*, Numer. Math., 3 (1961), pp. 374–380.

[39] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, *Matrix Eigensystem Routines*—EISPACK *Guide*, Lecture Notes in Computer Science Vol. 6, Springer-Verlag, New York, 1976.

[40] G. W. Stewart, *A Jacobi-like algorithm for computing the Schur decomposition of a non-Hermitian matrix*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 853–864.

[41] H. P. M. van Kempen, *On the convergence of the classical Jacobi method for real symmetric matrices with non-distinct eigenvalues*, Numer. Math., 9 (1966), pp. 11–18.

[42] K. Veselić, *On a new class of elementary matrices*, Numer. Math., 33 (1979), pp. 173–180.

[43] ———, *An Eigenreduction Algorithm for Definite Matrix Pairs and its Applications to Overdamped Linear Systems*, Tech. Report, Fakultät für Mathematik, Fernuniversität, Hagen, Germany, 1988.

[44] K. Veselić and H. J. Wenzel, *A quadratically convergent Jacobi-like method for real matrices with complex eigenvalues*, Numer. Math., 33 (1979), pp. 425–435.

[45] V. Voevodin, *An extension of the method of Jacobi*, in Computational Methods and Programming VIII, Computing Center of the Moscow University, Moscow, USSR, 1967, pp. 216–228. (In Russian.)

[46] J. H. Wilkinson, *Note on the quadratic convergence of the cyclic Jacobi process*, Numer. Math., 4 (1962), pp. 296–300.

[47] ———, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.

[48] ———, *Almost Diagonal Matrices with Multiple or Close Eigenvalues*, Tech. Report CS 59, Department of Computer Science, Stanford University, Stanford, CA, 1967.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 950–966, October 1993

003

ON THE CONVERGENCE OF REVERSIBLE MARKOV CHAINS*

MADHAV P. DESAI† AND VASANT B. RAO‡

Abstract. A simple upper bound for the second-largest eigenvalue of a finite reversible time-homogeneous Markov chain is presented as a function of the transition probabilities, the equilibrium distribution, and the underlying structure of the chain. This work is extended to a relation between the second-largest eigenvalues of any two reversible Markov chains with the same underlying structure. Furthermore, a lower bound for the smallest eigenvalue of a reversible chain is also presented, thereby providing a bound on the spectral gap of such chains. These eigenvalue bounds are fairly easy to compute for a variety of reversible chains by using known results on eigenvalues of certain matrices associated with graphs or random walks on graphs. The results on the spectral gap lead to a bound on the time constant of a reversible Markov chain converging to its equilibrium distribution. As an application, the temperature asymptotics of simulated annealing, which is a probabilistic algorithm widely used for solving combinatorial optimization problems, are studied.

Key words. reversible Markov chains, eigenvalue, underlying graph, Laplacian matrix, simulated annealing, temperature asymptotics, combinatorial optimization

AMS subject classifications. 60J10, 15A42

1. Introduction.

Let $\Omega = \{1, 2, \ldots, N\}$ be a discrete state space, and consider a time-homogeneous Markov chain $\{X(k)\}$ on Ω with an $N \times N$ probability transition matrix $P = [p_{ij}]$, where

$$p_{ij} = \text{Prob}\,(X(k+1) = j \,|\, X(k) = i) \tag{1.1}$$

for any pair of states $i, j \in \Omega$ and any time $k \geqq 0$. Let $v(k) = [v_i(k)]$ be the $1 \times N$ distribution vector describing the chain at time k such that $v_i(k) = \text{Prob}\,(X(k) = i)$ for all states $i \in \Omega$. It follows that $v(k+1) = v(k)P$. Suppose the Markov chain converges to an equilibrium distribution vector π, i.e.,

$$\lim_{k \to \infty} v(k) = \pi = \pi P. \tag{1.2}$$

In this paper we are primarily interested in the speed of convergence of $v(k)$ to π. If the Markov chain is reversible,[1] then the eigenvalues of P are real and are given as $1 = \lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_N \geqq -1$. It is then known [10], [21] that the error at time k can be bounded by

$$\|v(k) - \pi\| \leqq \pi_{\min}^{-1/2} \beta^k, \tag{1.3}$$

where $\beta = \max\,(\lambda_2, |\lambda_N|)$, $\|v\| = \sum_{i=1}^N |v_i|$ is the conventional 1-norm, and $\pi_{\min} = \min_i \pi_i$. The quantity $1 - \beta$ is often referred to as the *spectral gap* of the chain. Define

$$\tau = \frac{-1}{\log \beta} \tag{1.4}$$

to be the *time constant* of the Markov chain converging to its equilibrium distribution.

* Received by the editors June 25, 1990; accepted for publication (in revised form) December 31, 1991. This work was supported in part by Semiconductor Research Corporation contract SRC 86-12-109 and by U.S. Air Force Office of Scientific Research contract AFOSR 88-0181. This work was done while the authors were at the Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign.

† Digital Equipment Corporation, Hudson, Massachusetts 01749 (desai@decsim.enet.dec.com).

‡ IBM East Fishkill, Hopewell Junction, New York 12533 (raov@fshvmfk1.vnet.ibm.com).

[1] See Definition 2.1 for a definition of reversibility.

It follows that if $\beta \leqq 1 - 1/q$ for some $q > 1$, then $\tau \leqq q$. Furthermore, given any $0 < \delta < 1$ we will have the error $\|v(k) - \pi\| \leqq \delta$ for all times $k \geqq -(\log \delta + \frac{1}{2} \log \pi_{\min})\tau$. In this case we say that the Markov chain is within a distance δ from equilibrium. Therefore, the rate at which the Markov chain achieves equilibrium is determined by the time constant τ, which in turn depends on the spectral gap of the chain. An upper bound for τ has many applications in the study of probabilistic algorithms based on simulating Markov chains. For example, random walks may be used to simulate uniform distributions on a given set of objects [1]. In such cases, if τ is obtained as a function of the size of the state space, the computational complexity of the algorithm may be studied. Another application is a probabilistic search algorithm, such as simulated annealing (SA) [15]. In this algorithm the transition probabilities are controlled by a temperature parameter. For the SA algorithm we would require a bound for τ in terms of the size of the chain as well as the temperature. A more detailed discussion of such an application is presented in § 4.

One of the main results of this paper is the derivation of an upper bound on the second-largest eigenvalue of a reversible Markov chain. There is a considerable body of research in this area. Many successful approaches are based on the work of Cheeger [7], who considered the problem of finding a lower bound for the smallest positive eigenvalue of the Laplacian of a compact Riemannian manifold. Knowledge of this eigenvalue leads to bounds on the speed of convergence of the random process describing Brownian motion of a particle on the manifold. A discrete analogue of Cheeger's work was developed by Alon [2], who established the relationship between the smallest positive eigenvalue of the Laplacian matrix of a graph and a measure of its expansion properties. Alon's bound has the advantage of being easily computable, but direct application of his ideas to Markov chains leads to a useful bound only for random walks, as shown in [1]. Cheeger's work was generalized to reversible Markov processes on a general state space by Lawler and Sokal [16]. In the present paper we seek a useful bound for arbitrary discrete-time finite reversible Markov chains.

Sinclair and Jerrum [22] have derived a bound of the form $\lambda_2 \leqq 1 - \phi^2/2$, where ϕ is the conductance associated with a reversible Markov chain. The conductance is an extension of the expansion idea for edge-weighted graphs. The main effort of bounding λ_2 is then transferred to the task of bounding ϕ from below. Ingenious path-counting arguments have been used in [22] to get excellent bounds for some complex Markov chains. Another path-counting approach based on a Poincaré-type inequality was developed by Diaconis and Stroock [10]. This method gives results comparable to those in [22], but the task of calculating the bounds is still difficult, even for random walks on graphs. The difficulty involved in computing these bounds for arbitrary reversible chains (that are not random walks) is expected to be even greater.

The bound on $\lambda_2(P)$ derived in this paper is of the form $\lambda_2 \leqq 1 - \sigma\mu_2(Q)$, where σ is a simple function of the entries in P and its equilibrium distribution vector π, and $\mu_2(Q)$ is the second-smallest eigenvalue of Laplacian matrix Q associated with the underlying graph of P. Similarly, a lower bound for $\lambda_N(P)$ of the form $\lambda_N \geqq -1 + \sigma\mu_1(|Q|) + 2\gamma$ is also derived, where σ and γ are, once again, simple functions of the entries in P and the equilibrium distribution π, and $\mu_1(|Q|)$ is the smallest eigenvalue of the matrix $|Q|$ obtained by taking the absolute value of the entries of Q.

The special feature of the bounds derived in this paper is the clear separation of the effect of the underlying structure of P captured by $\mu_2(Q)$ (or $\mu_1(|Q|)$) and the effect of the entries in P and π captured by σ. Our bounds will be useful in the case of reversible chains for which $\mu_2(Q)$ is known exactly (e.g., cycles and hypercubes) or can be effectively bounded (e.g., graphs in [1]–[4], [9], [19]). In other cases one could bound $\mu_2(Q)$ by using known results on eigenvalues of random walks on graphs (e.g., graphs in [10], [22]). We also provide examples of reversible Markov chains (which are not random walks on graphs) for which our eigenvalue bounds are tight. As a penalty for the simplicity of their computation, there are instances in which our bounds may be pessimistic when compared with those obtained by more sophisticated methods, such as those in [10] and [22]. Our bounds, however, can provide first-order approximations to the speed of convergence of reversible chains in terms of their underlying structures and equilibrium distributions. By methods similar to those used in deriving the bounds on λ_2, we obtain a relation between the second-largest eigenvalues of two reversible chains that are structurally equivalent (i.e., that have the same underlying graph). This result is more general in the sense that knowledge of λ_2 for a reversible chain enables us to bound λ_2 for any structurally equivalent chain without explicitly considering the underlying graph.

The rest of this paper is organized as follows. In § 2 we establish some preliminaries. A new upper bound for the second-largest eigenvalue and a lower bound for the smallest eigenvalue of a reversible transition matrix are presented in § 3. In § 4 we briefly describe the simulated annealing algorithm and we discuss the application of our bounds to the corresponding reversible Markov chains.

2. Preliminaries and definitions. Consider a time-homogeneous Markov chain $\{X(k)\}$ on a finite state space $\Omega = \{1, 2, \ldots, N\}$ with transition matrix $P = [p_{ij}]$ as in § 1. We begin this section by reviewing some basic material on nonnegative matrices in general. In this paper we use standard graph-theoretic terminology from [6].

The *underlying directed graph* of P (or any matrix that is not necessarily a transition matrix) is a directed graph $G_d(\Omega, E_d)$ with vertex set Ω and an arc (i, j) directed from vertex i to vertex j if and only if $p_{ij} \neq 0$. The matrix P is *irreducible* if there exists a directed path from each vertex to every other vertex in its underlying directed graph G_d. For an irreducible matrix let r denote the greatest common divisor of the lengths of all the directed cycles in its underlying directed graph. If $r = 1$, the matrix is said to be *aperiodic*.

The Markov chain itself is said to be irreducible (aperiodic) if its transition matrix P is irreducible (aperiodic). Some basic facts on transition matrices [21] are as follows. Every eigenvalue of a transition matrix has magnitude $\leqq 1$. If P is irreducible, then 1 is a simple eigenvalue of P. Let π be the left eigenvector corresponding to eigenvalue 1 of an irreducible P, chosen so that $\sum_{i \in \Omega} \pi_i = 1$. Then $\pi_i > 0$ for each $i \in \Omega$ and π is termed the *equilibrium distribution vector*. An aperiodic Markov chain converges to its equilibrium distribution π for any initial distribution $v(0)$. The main results of this paper deal with reversible Markov chains, which are defined as follows.

DEFINITION 2.1. An irreducible Markov chain with transition matrix P and equilibrium distribution vector π is said to be *reversible* if for all $i, j \in \Omega$ we have

$$\pi_i p_{ij} = \pi_j p_{ji}. \tag{2.1}$$

From Definition 2.1 we see that a reversible Markov chain with transition matrix P is *structurally symmetric*, i.e., $p_{ij} > 0$, if and only if $p_{ji} > 0$. The *underlying undirected graph* of P (or any structurally symmetric matrix) is a simple undirected graph $G(\Omega, E)$ obtained from the underlying directed graph $G_d(\Omega, E_d)$ by deleting all self-loops and replacing directed 2-cycles by simple edges. Thus arcs (i, j) and (j, i) in G_d are replaced by a single edge $\{i, j\}$ in G.

A reversible Markov chain has the following interesting property. The proof is an easy consequence of Definition 2.1 and is therefore omitted.

PROPOSITION 2.2. *Consider a reversible Markov chain with transition matrix P and equilibrium distribution vector π. Define $d_i = \sqrt{\pi_i}$ for each $i \in \Omega$, and define the diagonal* *matrix*

$$D = \operatorname{diag}[d_1, d_2, \dots, d_N]. \tag{2.2}$$

Then the following hold:

(1) *D^2P is a symmetric matrix.*

(2) *DPD^{-1} is a symmetric matrix.*

(3) *P is diagonalizable, has real eigenvalues, and has a full set of linearly independent eigenvectors. If x and y are two linearly independent right eigenvectors of P, then they are D^2-orthogonal, i.e., $x^TD^2y = 0$.*

In general, for any $K \times K$ matrix M with real eigenvalues, let $\lambda_1(M) \geqq \lambda_2(M) \geqq \cdots \geqq \lambda_K(M)$ denote the eigenvalues of M arranged in descending order. Then for a transition matrix P of a reversible Markov chain we have

$$1 = \lambda_1(P) > \lambda_2(P) \geqq \lambda_3(P) \geqq \cdots \geqq \lambda_N(P) \geqq -1. \tag{2.3}$$

There are several symmetric matrices associated with undirected graphs. For this paper it suffices to consider only one of them.

DEFINITION 2.3. Consider a simple undirected graph $G(\Omega, E)$ on N vertices. Let $\deg(i)$ denote the degree of vertex $i \in \Omega$, which is the total number of edges in E incident on the vertex i. Then the *Laplacian matrix* $Q(G)$ is an $N \times N$ matrix with entries defined as

$$q_{ij} = \begin{cases} \deg(i) & \text{if } j = i, \\ -1 & \text{if } \{i, j\} \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

Remark. The Laplacian matrix Q used here is associated with the *adjacency matrix* [6] of the graph G. Some authors (see [10], for example) refer to $L = I - P$ as the Laplacian matrix associated with a transition matrix P. Relationships between the two Laplacians are discussed in Example 3.4 for the case in which P is a random walk on the graph G.

Clearly, the Laplacian matrix $Q(G)$ is a symmetric matrix. The following theorem provides some more information about $Q(G)$. The proof is an easy consequence of the discussion in [5], [6], [9] on spectra of graphs and is therefore omitted.

THEOREM 2.4. *If $G(\Omega, E)$ is a connected simple graph with Laplacian matrix Q, then the following hold*:

(1) *$Q\mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ is a vector with each entry $= 1$. Hence Q has an eigenvalue 0 with eigenvector $\mathbf{1}$. Moreover, 0 is a simple eigenvalue of Q, i.e.,* $\operatorname{rank}(Q) = N - 1$.

(2) *The quadratic form x^TQx equals $\sum_{\{i,j\}\in E}(x_i - x_j)^2$.*

In general, for any $K \times K$ matrix M with real eigenvalues let $\mu_1(M) \leqq \mu_2(M) \leqq \cdots \leqq \mu_K(M)$ denote the eigenvalues of M arranged in ascending order. From Theorem 2.4, Q is positive semidefinite and has eigenvalues

$$0 = \mu_1(Q) < \mu_2(Q) \leqq \cdots \leqq \mu_N(Q). \tag{2.5}$$

The following results are key to deriving the eigenvalue bounds in the next section.

LEMMA 2.5 (min-max principle).

(1) *If A is symmetric and z is the eigenvector corresponding to $\mu_1(A)$, then*

$$\min\left\{\frac{x^TAx}{x^Tx} : x \neq 0,\ z^Tx = 0\right\} = \mu_2(A). \tag{2.6}$$

(2) *If A and B are any two symmetric $K \times K$ matrices such that $A - B$ is positive semidefinite, then $\mu_i(A) \geqq \mu_i(B)$ for each $i = 1, 2, \ldots, K$.*

LEMMA 2.6. *Let B be any $(N-1) \times N$ matrix of full row rank. Then $\mu_i(BB^T) = \mu_{i+1}(B^TB)$ for each $i = 1, 2, \ldots, N-1$.*

The above two lemmas are well-known results from linear algebra. The proofs can be found in [14], for example.

THEOREM 2.7. *Let Q be any $N \times N$ symmetric and positive semidefinite matrix with* rank $(Q) = N - 1$, *and let D be an $N \times N$ diagonal matrix with strictly positive diagonal entries. Let $d_{\min} > 0$ denote the smallest diagonal entry in D. Then the following hold:*

(1) *There exists an $(N-1) \times N$ matrix B of full rank such that $Q = B^TB$.*

(2) *The $N \times N$ matrix DQD is symmetric and positive semidefinite.*

(3) $\mu_2(DQD) \geqq d_{\min}^2 \mu_2(Q)$.

Proof. Statements (1) and (2) are standard results and can be found in most textbooks on linear algebra. To prove (3), use (1) to write $Q = B^TB$, where B is an $(N-1) \times N$ matrix of full rank. Define $C = BD$. Clearly, C is also of full rank since D is a diagonal matrix with strictly positive diagonal entries. Note that $CC^T = BD^2B^T$ and that the matrix $CC^T - d_{\min}^2 BB^T$ is positive semidefinite since the quadratic form

$$x^T(CC^T - d_{\min}^2 BB^T)x = \sum_{i=1}^{N} (d_{ii}^2 - d_{\min}^2)[B^Tx]_i^2$$

is nonnegative for any vector $x \in \mathbb{R}^{N-1}$. Also, $DQD = DB^TBD = C^TC$. Therefore,

(2.7)

$$\mu_2(DQD) = \mu_2(C^TC) = \mu_1(CC^T) \geqq d_{\min}^2 \mu_1(BB^T) = d_{\min}^2 \mu_2(B^TB) = d_{\min}^2 \mu_2(Q),$$

where the second and third equalities are from Lemma 2.6 and the inequality is from part (2) of Lemma 2.5. □

3. A new eigenvalue bound for reversible Markov chains.

A reversible Markov chain has a structurally symmetric transition matrix P, as noted in § 2. Its underlying undirected graph G, therefore, is both connected and simple. Moreover, P is irreducible and has real eigenvalues from Proposition 2.2. Hence the second-largest eigenvalue of P is $\lambda_2 < 1$. One of the main results of this paper is a tighter upper bound for λ_2, which is provided in § 3.1. Another result relating the λ_2's of two different reversible chains having the same underlying graph is also derived. In § 3.2 a lower bound for the smallest eigenvalue $\lambda_N(P)$ is presented.

The following quantities defined for a transition matrix P with equilibrium distribution π will be used frequently:

(1) $w_{\min} = \min \{\pi_i p_{ij} : i \neq j, p_{ij} > 0\}$.

(2) $\pi_{\max} = \max_i \pi_i$.

(3) $\mu_2(Q) =$ the second-smallest eigenvalue of the Laplacian matrix Q of the underlying undirected graph G of P.

3.1. An upper bound for λ_2.

THEOREM 3.1. *Consider a reversible Markov chain on the state space $\Omega = \{1, 2, \ldots, N\}$, with transition matrix P and equilibrium distribution π. Let $w_{\min}$, $\pi_{\max}$, and $\mu_2(Q)$ be as defined above. If $\lambda < 1$ is an eigenvalue of P, then*

$$\lambda \leqq 1 - \frac{w_{\min}}{\pi_{\max}} \mu_2(Q). \tag{3.1}$$

Proof. Let $d_i = \sqrt{\pi_i}$ for each $i \in \Omega$, and define the $N \times N$ diagonal matrix $D = \operatorname{diag}[d_1, d_2, \ldots, d_N]$. Since P is irreducible, we have that D is invertible and $D^{-1} = \operatorname{diag}[d_1^{-1}, d_2^{-1}, \ldots, d_N^{-1}]$.

Let $\lambda < 1$ be an eigenvalue of P, and let $x \in \mathbb{R}^N$ be the corresponding right eigenvector, i.e., $Px = \lambda x$. Note that $x^T D^2 (I - P)x = (1 - \lambda)x^T D^2 x$, which can be written as

$$1 - \lambda = \frac{x^T(D^2 - W)x}{x^T D^2 x}, \tag{3.2}$$

where we have defined the matrix $W = D^2 P$, which is symmetric by Proposition 2.2, part (1). For each $i = 1, 2, \ldots, N$ we have

$$\sum_{j=1}^{N} w_{ij} = \pi_i \sum_{j=1}^{N} p_{ij} = \pi_i. \tag{3.3}$$

Now consider the quadratic form in the numerator of (3.2), which can be written as

$$x^T(D^2 - W)x = \sum_{i=1}^{N} (\pi_i - w_{ii})x_i^2 - \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij} x_i x_j. \tag{3.4}$$

Using (3.3), we get

$$x^T(D^2 - W)x = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij}(x_i^2 - x_i x_j). \tag{3.5}$$

Now consider $G(\Omega, E)$, the underlying undirected graph of P. This is also the underlying graph of the symmetric matrix W, since for $j \neq i$ we have $p_{ij} \neq 0$ if and only if $w_{ij} \neq 0$. Hence (3.5) can be written as

$$x^T(D^2 - W)x = \sum_{\{i,j\} \in E} w_{ij}(x_i - x_j)^2 \geqq w_{\min} \sum_{\{i,j\} \in E} (x_i - x_j)^2. \tag{3.6}$$

Applying Theorem 2.4, part (2), to the right-hand side of (3.6) results in

$$x^T(D^2 - W)x \geqq w_{\min} x^T Q x, \tag{3.7}$$

where Q is the Laplacian matrix associated with the underlying graph G. Combining (3.2) and (3.7) gives

$$1 - \lambda \geqq w_{\min} \frac{x^T Q x}{x^T D^2 x} = w_{\min} \frac{y^T D^{-1} Q D^{-1} y}{y^T y}, \tag{3.8}$$

where we have defined $y = Dx$. Since $P\mathbf{1} = \mathbf{1}$, we have $x^T D^2 \mathbf{1} = 0$ from Proposition 2.2, part (3). Hence $y^T D\mathbf{1} = x^T D^2 \mathbf{1} = 0$ and $y \neq 0$. Also, $Q\mathbf{1} = \mathbf{0}$, which implies that $D\mathbf{1}$ is an eigenvector of $D^{-1}QD^{-1}$ corresponding to $\mu_1(D^{-1}QD^{-1}) = 0$. Therefore, from (2.6) in Lemma 2.5 we have

$$\frac{y^T D^{-1} Q D^{-1} y}{y^T y} \geqq \mu_2(D^{-1}QD^{-1}). \tag{3.9}$$

Applying Theorem 2.7, part (3), to the right-hand side of (3.9) gives

$$\mu_2(D^{-1}QD^{-1}) \geqq \frac{1}{\pi_{\max}} \mu_2(Q). \tag{3.10}$$

Finally, combining (3.8), (3.9), and (3.10) proves that (3.1) is true and hence proves the theorem. □

Remark. Note that (3.1) may be rewritten as

$$\mu_2(I-P) \geqq \frac{w_{\min}}{\pi_{\max}} \mu_2(Q). \tag{3.11}$$

This was derived by using the min–max principle (Lemma 2.5) and the fact that **1** is a common right eigenvector to both $(I-P)$ and Q corresponding to eigenvalues $\mu_1(I-P)=0$ and $\mu_1(Q)=0$, respectively. It might seem that a relation similar to (3.11) between $\mu_i(I-P)$ and $\mu_i(Q)$ for $i>2$ could be proved by extending the min–max principle for other eigenvalues. However, this is not possible since the smallest $(i-1)$ eigenvalues of $(I-P)$ and Q may not in general share a common set of eigenvectors for $i>2$.

For some graphs G the second-smallest eigenvalue $\mu_2(Q(G))$ is easy to compute analytically. Two examples are given below.

Example 3.2 (cycle graphs). If G is a simple cycle on N vertices, then the eigenvalues of its Laplacian matrix Q can be shown to be (e.g., see [5], [9])

$$2(1-\cos(2\pi(i-1)/N)) \quad \text{for } i=1,2,\dots,N. \tag{3.12}$$

Consequently, $\mu_2(Q)=2(1-\cos(2\pi/N))$, which approaches 0 as $N \to \infty$.

Example 3.3 (hypercube graphs). If G is an n-dimensional hypercube having $N=2^n$ vertices, then its Laplacian matrix Q has $n+1$ distinct eigenvalues given by (e.g., see [9], [10])

$$\xi_m = 2m \quad \text{for } m=0,1,2,\dots,n. \tag{3.13}$$

The eigenvalue $2m$ has an algebraic multiplicity $\binom{n}{m}$. Consequently, the second-smallest eigenvalue $\mu_2(Q)=2$ is independent of N, the size of the matrix.

Remark. For graphs G for which $\mu_2(Q(G))$ is not easy to compute exactly, one can use a lower bound. For example, Alon [2] relates $\mu_2(Q(G))$ to the expansion properties of the graph G. Another class of graphs for which $\mu_2(Q)$ has been estimated is the class of *Cayley graphs* of a finite group [1], [4]. There has been considerable related work on finding $\mu_2(Q)$ for a variety of graphs, as summarized in [9], [19]. Alternatively, one could relate $\mu_2(Q)$ to the known results on eigenvalues of random walks on graphs as follows.

Example 3.4 (random walk on a graph). Let $G=(\Omega,E)$ be a connected simple graph with minimum degree d_* and maximum degree d^*. Let P be the transition matrix of a random walk on G with entries

$$p_{ij} = \begin{cases} [\deg(i)]^{-1} & \text{if } \{i,j\} \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

Then P has underlying graph G and is reversible with respect to the equilibrium vector π with entries $\pi_i = \deg(i)/(2|E|)$. It follows that $w_{\min}=1/(2|E|)$ and that $\pi_{\max} = d^*/(2|E|)$. Then Theorem 3.1 gives

$$\lambda_2(P) \leqq 1 - \frac{1}{d^*}\mu_2(Q(G)). \tag{3.15}$$

Application of Alon's lower bound [2] for $\mu_2(Q)$ yields the same upper bound for $\lambda_2(P)$ as the one derived in [1]. For regular graphs (i.e., graphs for which $d_* = d^*$), the inequality in (3.15) is, in fact, an equality. The graphs in Examples 3.2 and 3.3 are regular graphs for which (3.15) provides $\lambda_2(P)$ exactly. Recently there has been a flurry of activity in lower-bounding $1-\lambda_2(P)=\mu_2(L)$ for random walks, where $L=I-P$

is called the Laplacian matrix associated with the transition matrix P. It is clear that the two Laplacian matrices are related by $Q = \hat{D}L$, where $\hat{D}$ is a diagonal matrix with ith diagonal entry $= \deg(i)$. Hence

$$d^*\mu_2(L) = d^*\mu_2(\hat{D}^{-1/2}Q\hat{D}^{-1/2}) \geqq \mu_2(Q) = \mu_2(\hat{D}^{1/2}L\hat{D}^{1/2}) \geqq d_*\mu_2(L), \tag{3.16}$$

where the equalities are from similarity transformations and the inequalities follow from Theorem 2.7, part (3). One could then use (3.16) and lower bounds for $\mu_2(L)$ from sources such as [9], [10], [16] to get lower bounds for $\mu_2(Q)$. A recent paper by Diaconis and Stroock [10] summarizes much of the activity in obtaining lower bounds for $\mu_2(L)$ for random walks. For regular graphs the inequalities in (3.16) become equalities.

It must be emphasized that the utility of Theorem 3.1 is not in obtaining bounds for $\lambda_2(P)$ of random walks on graphs but rather in bounding $\lambda_2(P)$ of any reversible Markov chain as a simple (and easily computable) function of the entries of P and π and the structure of the chain as captured through $\mu_2(Q)$.

We now relate $\lambda_2(P)$ and $\lambda_2(P')$ of any two reversible Markov chains $P = [p_{ij}]$ and $P' = [p'_{ij}]$ having the same underlying graph $G(\Omega, E)$. To this end, let quantities associated with P' be denoted as π', W', D', etc., analogous to similar quantities associated with P, as in Theorem 3.1.

THEOREM 3.5.[2] *Let P and P' be the transition matrices of two reversible chains with a common underlying graph $G(\Omega, E)$. Then*

$$1 - \lambda_2(P) \geqq \frac{\min_{\{i,j\}\in E}(w_{ij}/w'_{ij})}{\max_i(\pi_i/\pi'_i)}(1 - \lambda_2(P')). \tag{3.17}$$

Proof. The proof is similar to that of Theorem 3.1. We provide only an outline here and omit the details for the sake of brevity. Let x be the right eigenvector of P corresponding to $\lambda_2(P)$, and define $z = x - (\sum_{i=1}^{N}\pi'_i x_i)\mathbf{1}$. Since x and $\mathbf{1}$ are linearly independent and D^2-orthogonal, we have $z \neq \mathbf{0}$ and $z^TD^2z \geqq x^TD^2x$. Moreover, $z^T(D^2 - W)z = x^T(D^2 - W)x$ since $(D^2 - W)\mathbf{1} = \mathbf{0}$. By proceeding as in the proof of Theorem 3.1 it can be shown that

$$\begin{aligned} 1 - \lambda_2(P) &= \frac{x^T(D^2 - W)x}{x^TD^2x} \\ &\geqq \frac{z^T(D^2 - W)z}{z^TD^2z} \\ &\geqq \frac{\min_{\{i,j\}\in E}(w_{ij}/w'_{ij})}{\max_i(\pi_i/\pi'_i)}\frac{z^T(D'^2 - W')z}{z^TD'^2z} \\ &\geqq \frac{\min_{\{i,j\}\in E}(w_{ij}/w'_{ij})}{\max_i(\pi_i/\pi'_i)}\mu_2(I - D'P'D'^{-1}) \\ &= \frac{\min_{\{i,j\}\in E}(w_{ij}/w'_{ij})}{\max_i(\pi_i/\pi'_i)}(1 - \lambda_2(P')), \end{aligned} \tag{3.18}$$

which completes the proof. □

Remark. In Theorem 3.1 we related the eigenvalue of a reversible chain to its underlying structure. The result of Theorem 3.5 extends this line of thought a step further by directly relating $\lambda_2(P)$ of a reversible chain P with $\lambda_2(P')$ of any other reversible chain

[2] This result was provided by one of the referees of this paper.

P' that has the same underlying graph, without explicitly computing the $\mu_2(Q)$ of the underlying graph. This is especially useful in the case for which $\lambda_2(P')$ can be computed exactly or at least can be bounded tightly. Such is the case when P' is a random walk on a graph [9], [10], [16].

3.2. A lower bound for λ_N. The rate of convergence of a reversible Markov chain is governed by the spectral gap $1 - \beta = \min(1 - \lambda_2, 1 + \lambda_N)$, as seen in § 1. In § 3.1 we provided bounds for $1 - \lambda_2$. In this section we present a simple lower bound for $1 + \lambda_N$ of a reversible P as a function of the entries in P and its underlying structure. The two bounds together will provide a bound on the spectral gap. Before stating the main result, we define a few terms. Consider the reversible Markov chain P defined in the statement of Theorem 3.1. Let $G(\Omega, E)$ be the underlying graph of P. Define the matrix $|Q|$ with ijth entry $|q_{ij}|$, where q_{ij}'s are the entries of the Laplacian matrix Q of G as defined by (2.4). Let $\mu_1(|Q|)$ denote the smallest eigenvalue of $|Q|$. From (3.24) below we shall see that $|Q|$ is nonnegative definite; hence $\mu_1(|Q|) \geqq 0$. In addition, let $\gamma = \min_i p_{ii}$ denote the smallest diagonal entry in P. The rest of the terminology is the same as that in Theorem 3.1.

THEOREM 3.6. *The smallest eigenvalue λ_N of P satisfies*

$$\lambda_N \geqq -1 + \frac{w_{\min}}{\pi_{\max}} \mu_1(|Q|) + 2\gamma. \tag{3.19}$$

If the underlying graph G is k-regular (i.e., the degree of each vertex is k), then

$$\mu_1(|Q|) = 2k - \mu_N(Q), \tag{3.20}$$

where $\mu_N(Q)$ is the largest eigenvalue of the Laplacian matrix Q of G.

Proof. This proof uses the same notation as that of Theorem 3.1. Note that $1 + \lambda_N$ is the smallest eigenvalue of the matrix $I + P$. Let x be the corresponding right eigenvector. Then, as in the proof of Theorem 3.1, we can write

$$1 + \lambda_N = \frac{x^T D^2 (I + P) x}{x^T D^2 x}. \tag{3.21}$$

By proceeding as in the proof of Theorem 3.1, it can be verified that

$$x^T D^2 (I + P) x = 2 \sum_{i=1}^{N} \pi_i p_{ii} x_i^2 + \sum_{\{i,j\} \in E} \pi_i p_{ij} (x_i + x_j)^2. \tag{3.22}$$

Using $\gamma = \min_i p_{ii}$ and $\pi_i p_{ij} = w_{ij} \geqq w_{\min}$, we get

$$x^T D^2 (I + P) x \geqq 2\gamma \sum_{i=1}^{N} \pi_i x_i^2 + w_{\min} \sum_{\{i,j\} \in E} (x_i + x_j)^2. \tag{3.23}$$

Analogously to Theorem 2.4, part (2), we can write

$$\sum_{\{i,j\} \in E} (x_i + x_j)^2 = x^T |Q| x. \tag{3.24}$$

Since $|Q|$ is symmetric, we have $x^T |Q| x \geqq \mu_1(|Q|) x^T x$. Hence

$$x^T D^2 (I + P) x \geqq 2\gamma x^T D^2 x + w_{\min} \mu_1(|Q|) x^T x. \tag{3.25}$$

But $x^T D^2 x \leqq \pi_{\max} x^T x$. Combining this with (3.21) and (3.25) proves (3.19). To prove (3.20) note that if G is k-regular, then $|Q| = 2kI - Q$. □

Remark. Suppose $P = [p_{ij}]$ and $P' = [p'_{ij}]$ are the $N \times N$ transition matrices of two reversible chains with the same underlying *directed* graph $G_d(\Omega, E_d)$. For two such matrices we can obtain a result analogous to Theorem 3.5 for their smallest eigenvalues. By using the same terminology as in Theorem 3.5, it is easy to show that

$$1 + \lambda_N(P) \geqq \frac{\min_{(i,j)\in E_d}(w_{ij}/w'_{ij})}{\max_{i\in\Omega}(\pi_i/\pi'_i)}(1 + \lambda_N(P')).$$

In conjunction with Theorem 3.5, this provides a relationship between the spectral gaps of two reversible Markov chains with the same underlying directed graph.

Example 3.7 (a reversible chain that is not a random walk). Consider the following reversible Markov chain on the vertex set $\{1, 2, \ldots, N\}$. Let $0 < \varepsilon < 1$, and define the transition matrix P by $p_{ij} = 1/(N-1)$ for $i = 2, 3, \ldots, N$ and $j \neq i$, and by $p_{1j} = \varepsilon/(N-1)$ for $j \neq 1$. The diagonal entries of P are $p_{11} = 1 - \varepsilon$ and $p_{ii} = 0$ for $i > 1$. The underlying graph of P is the complete graph on N vertices, which is $(N-1)$-regular and is denoted by K_N. Let $Z = 1 + (N-1)\varepsilon$. Then the equilibrium distribution is given by $\pi_1 = 1/Z$ and $\pi_i = \varepsilon/Z$ for $i > 1$. Hence $w_{ij} = \varepsilon/(Z(N-1))$ for all $i \neq j$, $\pi_{\max} = 1/Z$, and $\gamma = 0$. If Q is the Laplacian matrix of K_N, then $\mu_2(Q) = \mu_N(Q) = N$. Therefore, $\mu_1(|Q|) = N - 2$ from (3.20). The exact eigenvalues[3] of P for this example are

$$\lambda_1(P) = 1 \tag{3.26}$$

with $\mathbf{1}$ as the corresponding right eigenvector,

$$\lambda_2(P) = \frac{N-2}{N-1} - \varepsilon = 1 - \frac{Z}{N-1} \tag{3.27}$$

with the entries of the corresponding right eigenvector defined as $x_1 = -\varepsilon(N-1)$ and $x_2 = x_3 = \cdots = x_N = 1$, and, finally,

$$\lambda_3(P) = \lambda_4(P) = \cdots = \lambda_N(P) = \frac{-1}{N-1} \tag{3.28}$$

with the corresponding $(N-2)$-dimensional right eigenspace defined as $\{y \in \mathbb{R}^N : y_1 = 0, \sum_{i=2}^{N} y_i = 0\}$. The results of § 3 can be used to bound $\lambda_2(P)$ and $\lambda_N(P)$ as follows. Using Theorem 3.1, we get

$$\lambda_2(P) \leqq 1 - \varepsilon\frac{N}{N-1}, \tag{3.29}$$

and, using Theorem 3.6, we get

$$\lambda_N(P) \geqq -1 + \varepsilon\frac{N-2}{N-1}. \tag{3.30}$$

If Theorem 3.5 is used with P' corresponding to the random walk on K_N, we get the same bound for $\lambda_2(P)$ as in (3.29) since in this case $\lambda_2(P') = -1/(N-1)$, $w'_{ij} = 1/(N(N-1))$ for all $i \neq j$, and $\pi'_i = 1/N$ for all i. A comparison of (3.27) and (3.29) shows that the upper bound is off by $(1-\varepsilon)/(N-1)$, and comparing (3.28) and (3.30) shows that the lower bound is off by $(1-\varepsilon)(N-2)/(N-1)$. From (3.29) and (3.30), the spectral gap of P can be bounded as

[3] The authors are very grateful to one of the referees of this paper for providing us with the exact eigenvalues for Example 3.7.

$$1-\beta \geqq \varepsilon \frac{N-2}{N-1}, \tag{3.31}$$

whereas the exact eigenvalues result in

$$1-\beta=\frac{\min\{Z,(N-2)\}}{N-1}. \tag{3.32}$$

Therefore, the bound of (3.31) is off by $(1+\varepsilon)/(N-1)$ from the exact value of the spectral gap for $0<\varepsilon \leqq (N-3)/(N-1)$, whereas (3.31) is off by $(1-\varepsilon)(N-2)/(N-1)$ for $(N-3)/(N-1)<\varepsilon<1$. In either case the bound on the spectral gap is tight for large N.

Remark. If a graph G is bipartite, then $\mu_1(|Q|)=0$ with the corresponding eigenvector x having entries $x_i=+1$ if vertex i is in one part and having entries $x_i=-1$ if i is in the other part. Hence if we have a reversible Markov chain P on such a graph, then Theorem 3.6 gives $\lambda_N \geqq -1+2\gamma$. If $p_{ii}=0$ for all i, then it can be shown that $\lambda_N=-1$, which agrees with the lower bound since $\gamma=0$ in this case. On the other hand, if $\gamma=0$ but $p_{ii}>0$ for some i, then P is aperiodic, which implies that $\lambda_N>-1$, whereas Theorem 3.6 simply gives $\lambda_N \geqq -1$. In such a situation one could use the results in [10], in which a lower bound for λ_N of a random walk is provided as follows. Let ζ_i denote a closed path from a vertex i to itself that contains an odd number of edges in the underlying graph $G(\Omega, E)$ of a random walk. Let Ψ be a set of n such paths in G, one for each vertex in Ω. Define the quantity

$$\theta(\Psi)=\max_{\{i,j\}\in E}\left(\sum_{\{\zeta_i\in\Psi:\{i,j\}\in\zeta_i\}} |\zeta_i|\pi_i\right).$$

Then, [10, Prop. 2] states that

$$\lambda_N \geqq -1+\frac{2}{\theta(\Psi)}.$$

Alternatively, one could use the approach in [23], where a lower bound for $\mu_1(|Q|)$ is derived as a function of a measure of the nonbipartiteness of the graph.

4. Applications of the eigenvalue bounds.

As an application of the results of § 3, we consider the simulated annealing algorithm, which is a probabilistic algorithm first proposed in [15] for solving difficult combinatorial optimization problems in the design of very-large-scale integrated (VLSI) circuits.

The SA algorithm can be briefly described as follows. Given a state space $\Omega=\{1,2,\ldots,N\}$ with a cost function $C:\Omega\rightarrow\mathbb{R}$, the SA algorithm attempts to find a state with minimum cost. The algorithm can be modeled by a reversible Markov chain whose underlying graph $G(\Omega, E)$ is connected. For simplicity of presentation we assume that G is $(\rho-1)$-regular, where $\rho\geqq 3$ is an integer. The transition probabilities are controlled by a parameter $T>0$ called the *temperature*. To further simplify notation we define

$$\varepsilon=e^{-1/T}. \tag{4.1}$$

We will henceforth refer to ε as the *temperature parameter*. Given $\varepsilon>0$, the SA algorithm is described by the transition matrix $P=[p_{ij}]$ whose off-diagonal entries $(i\neq j)$ are given by

$$p_{ij}=\begin{cases}\rho^{-1}\varepsilon^{[C(j)-C(i)]^+} & \text{if } \{i,j\}\in E,\\ 0 & \text{otherwise,}\end{cases} \tag{4.2}$$

where $[z]^+ = z$ if $z > 0$, and $[z]^+ = 0$ if $z \leqq 0$. The diagonal entries of P are given by

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}. \tag{4.3}$$

It follows from (4.2) and (4.3) that $p_{ii} \geqq 1/\rho$ for each $i \in \Omega$. This, combined with the fact that G is connected, implies that P is aperiodic. Furthermore, the equilibrium distribution vector $\pi(\varepsilon)$ can be shown (see [18]) to have entries

$$\pi_i(\varepsilon) = \frac{\varepsilon^{C(i)}}{\sum_{l=1}^{N} \varepsilon^{C(l)}}; \tag{4.4}$$

$\pi(\varepsilon)$ is called the *Boltzmann distribution* at temperature ε. The denominator of (4.4) is called the *partition function*, and we shall denote it by $Z(\varepsilon)$. Using (4.2) and (4.4), one can verify that the SA chain is reversible and thus converges to its equilibrium distribution. From the discussion in § 1 the rate of this convergence is governed by the time constant τ as defined by (1.4).

For the SA chain, $\lim_{\varepsilon \to 0} \pi(\varepsilon) = \pi^*$ [18], where π^* is the *optimal distribution* vector, which is zero outside the set of states with globally minimum cost and is uniformly distributed within this set. Several researchers have shown that if ε is allowed to vary with time k according to a prespecified cooling schedule, then the resulting time-inhomogeneous Markov chain converges to the optimal distribution (see [8], [12], [18] for details).

One may also hold ε fixed and use the SA algorithm as an approximation method as follows. Given a $\delta > 0$, suppose that a state $i \in \Omega$ of cost $C(i) < C_{\min} + \delta$ is desired, where $C_{\min} = \min_j C(j)$. For any arbitrarily small $\alpha > 0$ the temperature ε may be selected so that the equilibrium probability of the SA chain's being in the desired set of states is at least $1 - \alpha$. If the SA chain is simulated until it is almost at equilibrium, then a desired state will be obtained with probability at least $1 - \alpha$. The above experiment may be repeated k times to obtain a desired state with probability at least $1 - \alpha^k$, which can be made sufficiently close to 1 for large k. In the above approximation algorithm the determination of a suitable temperature ε is by no means an easy task, in general, since it involves a careful study of the distribution of costs among the states. However, this approach has been used successfully in some problems, such as finding maximum matchings in graphs [20], [22] and finding integer compositions from a given set of positive integers [17]. The computational time required for such an approximation algorithm is determined by the time constant τ of the SA Markov chain, which we will bound by using the results of § 3.

Temperature asymptotics of SA. By temperature asymptotics of SA we mean the behavior of the spectral gap $1 - \beta$ and time constant τ as $\varepsilon \to 0$. We will use Theorem 3.1 and Theorem 3.6 to obtain bounds for the spectral gap of the SA chain as a function of ε. From these results we can also bound the time required for the SA chain to approximate equilibrium at any temperature.

Let us now relate the parameters used in our bounds to the parameters of the optimization problem being solved by the SA Markov chain. Define $C_{\max} = \max_j C(j)$ and $\Delta = C_{\max} - C_{\min}$. If $Z(\varepsilon)$ is the partition function, then $w_{ij} = \varepsilon^{\max\{C(i), C(j)\}}/(\rho Z(\varepsilon))$ for each edge $\{i, j\} \in E$, and $\pi_i = \varepsilon^{C(i)}/Z(\varepsilon)$ for each state $i \in \Omega$. Therefore, $w_{\min}/\pi_{\max} = \varepsilon^{\Delta}/\rho$. Theorems 3.1 and 3.6 result in

$$\lambda_2(P) \leqq 1 - \frac{\mu_2(Q)\varepsilon^{\Delta}}{\rho} \quad \text{and} \quad \lambda_N(P) \geqq -1 + \frac{2}{\rho} \tag{4.5}$$

since $\gamma \geqq 1/\rho$ and $\mu_1(|Q|) \geqq 0$. From (4.5) and (1.4) the spectral gap and the time constant of the SA chain can be bounded as

$$(4.6) \qquad 1-\beta \geqq \frac{\mu_2(Q)\varepsilon^{\Delta}}{\rho} \quad \text{and} \quad \tau \leqq \frac{\rho}{\mu_2(Q)\varepsilon^{\Delta}}$$

when $\varepsilon \leqq [2/\mu_2(Q)]^{1/\Delta}$.

Remark. For a fixed optimization problem (i.e., fixed parameters ρ, Δ, $\mu_2(Q)$, etc.), (4.6) shows that the time constant of the SA Markov chain is $\tau = O(\varepsilon^{-\Delta})$. The recent work of Holley and Stroock [13] indicates that a better bound exists as $\varepsilon \to 0$. Their result is briefly summarized as follows. Let $\Xi_{i,j}$ denote the set of all paths between two vertices (states) i and j in the underlying graph G of the Markov chain. For each path $\xi \in \Xi_{i,j}$ define the height of the path to be $h(\xi) = \max_{l \in \xi} C(l)$. For the SA chain define the *depth parameter* $d = \max_{i,j \in \Omega} \min_{\xi \in \Xi_{i,j}} (h(\xi) - C(i))$. Clearly, d is bounded above by Δ, but it may be smaller in many instances. In [13] it is shown that as $\varepsilon \to 0$ the spectral gap is $1 - \beta = \Theta(\varepsilon^d)$ (i.e., there exist constants c_1 and c_2 such that $c_1\varepsilon^d \leqq 1 - \beta \leqq c_2\varepsilon^d$). This suggests that there is scope for improvement in the bounds of Theorems 3.1 and 3.6 when $d < \Delta$.

We will now consider the case of the SA chain operating under a prespecified cooling schedule. In particular, we show that knowledge of λ_2 at some temperature provides information about λ_2 at any other smaller temperature. Consider the cooling schedule [8], [12], [18] $\{\varepsilon_m\}$, which is typically of the form $\varepsilon_m = (m+1)^{-1/\Delta}$ for $m \geqq 0$. Note that the case $\varepsilon_0 = 1$ corresponds to the random walk on the graph G. Let P_m denote the transition matrix at temperature ε_m. Using Theorem 3.5, for the case $0 < \varepsilon_l < \varepsilon_m \leqq 1$ we see that $(\varepsilon_m/\varepsilon_l)^{\Delta}(1 - \lambda_2(P_m)) \geqq (1 - \lambda_2(P_l)) \geqq (\varepsilon_l/\varepsilon_m)^{\Delta}(1 - \lambda_2(P_m))$. For successive temperatures of the typical cooling schedule, this works out to

$$(4.7) \qquad \frac{m+2}{m+1}(1 - \lambda_2(P_m)) \geqq (1 - \lambda_2(P_{m+1})) \geqq \frac{m+1}{m+2}(1 - \lambda_2(P_m)).$$

For large m, (4.7) illustrates the variation of the spectral gap at successive temperatures for the typical cooling schedule.

We will next provide an example of a cost distribution on a state space for which the eigenvalue bound of Theorem 3.1 for the SA transition matrix is quite tight when treated as a function of temperature ε alone. We will also compare our bound with those of Sinclair and Jerrum [22] and of Diaconis and Stroock [10] for this example. The bound in [22] is based on an analogue of the expansion idea for an edge-weighted graph, which is briefly described as follows. Consider a reversible Markov chain on state space Ω with transition matrix P and equilibrium distribution π. The *conductance* of P is defined as

$$(4.8) \qquad \phi(P) = \min \frac{\sum_{\{i,j\} \in \mathrm{Cut}(S)} \pi_i p_{ij}}{\sum_{i \in S} \pi_i},$$

where the minimization is over all subsets S of states with $0 < \sum_{i \in S} \pi_i \leqq \frac{1}{2}$ and Cut (S) denotes the set of all edges $\{u, v\}$ in the underlying graph with $u \in S$ and $v \notin S$. The conductance ϕ is also called the *Cheeger constant*, and the set S at which the minimum of (4.8) is achieved is called the *Cheeger set*. It can be shown [22] that

$$(4.9) \qquad 1 - 2\phi(P) \leqq \lambda_2(P) \leqq 1 - \frac{\phi(P)^2}{2}.$$

An alternative lower bound for λ_2 was developed by Diaconis and Stroock on the basis of the Poincaré inequality [10]. Let P, π, W, and $G(\Omega, E)$ be as in Theorem 3.1. For each pair of distinct states i and j in Ω specify a *canonical path* $\xi(i,j)$ from i to j in the underlying graph G. Denote by $|\xi(i,j)|$ the number of edges in the canonical path from i to j. Let Γ denote the collection of canonical paths (one for each ordered pair $(i,j) \in \Omega \times \Omega$). Define the quantity

$$K(\Gamma) = \max_{\{i,j\}\in E} \left(\frac{1}{w_{ij}} \sum_{\xi(u,v)\ni\{i,j\}} |\xi(u,v)|\pi_u\pi_v \right), \tag{4.10}$$

where the inner summation is carried out over all canonical paths $\xi(u,v) \in \Gamma$ that contain the edge $\{i,j\}$. Then it can be shown [10] that

$$\lambda_2(P) \leqq 1 - \frac{1}{K(\Gamma)}. \tag{4.11}$$

A judicious choice of the set of canonical paths Γ is critical in developing a tight bound. Also, the conductance ϕ may be bounded by using the set of canonical paths Γ as follows. Define the quantity

$$U(\Gamma) = \max_{\{i,j\}\in E} \left(\frac{1}{w_{ij}} \sum_{\xi(u,v)\ni\{i,j\}} \pi_u\pi_v \right). \tag{4.12}$$

Then, from [10], $\phi \geqq 1/(2U(\Gamma))$ and also $K(\Gamma) \leqq l_{\max}U(\Gamma)$, where $l_{\max}$ is the length of the longest path in Γ. The quantity U in (4.12) is usually easier to compute than is K in (4.10) or ϕ in (4.8).

The following simple example will be used as a test case to compare the different eigenvalue bounds.

Example 4.1. Consider a simple cycle on $N = 4n$ vertices as the underlying graph of an SA Markov chain with a cost function defined as follows:

$$C(i) = \begin{cases} i & \text{if } 1 \leqq i \leqq n, \\ 2n+1-i & \text{if } n+1 \leqq i \leqq 2n, \\ i-2n & \text{if } 2n+1 \leqq i \leqq 3n, \\ 4n+1-i & \text{if } 3n+1 \leqq i \leqq 4n. \end{cases} \tag{4.13}$$

At a given $\varepsilon > 0$ the transition matrix P is obtained from (4.2) and (4.3) by using the costs from (4.13) and $\rho = 3$. The corresponding equilibrium distribution π can be found by using (4.4). For this SA chain we can obtain the following bounds for $\lambda_2(P)$. For transition matrix P it can be shown that $\pi_{\max} = (1-\varepsilon)/(4(1-\varepsilon^n))$, $w_{\min} = \varepsilon^{n-1}(1-\varepsilon)/(12(1-\varepsilon^n))$, and $\mu_2(Q) = 2(1-\cos(\pi/(2n)))$ because the underlying graph is a cycle on $4n$ vertices. Thus the bound from Theorem 3.1 gives

$$1 - \lambda_2(P) \geqq \frac{2}{3}\varepsilon^{n-1}\left(1 - \cos\left(\frac{\pi}{2n}\right)\right) \sim \frac{\pi^2}{12n^2}\varepsilon^{n-1}, \tag{4.14}$$

where the approximation holds for large n (uniformly in ε). The conductance of the chain in this example can be computed to be $\phi = \varepsilon^{n-1}(1-\varepsilon)/(3(1-\varepsilon^n))$ with $S = \{n, n+1, \ldots, 3n\}$ as a Cheeger set. Hence the Sinclair–Jerrum bounds from (4.9) imply that

$$\frac{2\varepsilon^{n-1}(1-\varepsilon)}{3(1-\varepsilon^n)} \geqq 1 - \lambda_2(P) \geqq \frac{\varepsilon^{2n-2}(1-\varepsilon)^2}{18(1-\varepsilon^n)^2}. \tag{4.15}$$

To compute the Diaconis–Stroock bound for this example we choose the shorter of the two paths from i to j on the cycle as the canonical path $\xi(i, j)$ for each pair of distinct states i and j in Ω. It is easy to see that $|\xi(i, j)| \leqq 2n$. The number of canonical paths containing a particular edge in the graph can be shown to be at most $1 + 2 + \cdots + 2n = n(2n + 1)$. With this information one can bound the quantity in (4.10) as $K \leqq 2n\pi_{\max}^2 n(2n + 1)/w_{\min} = 3n^2(2n + 1)(1 - \varepsilon)/(2\varepsilon^{n-1}(1 - \varepsilon^n))$. Hence from (4.11) we get the Diaconis–Stroock bound

$$(4.16) \qquad 1 - \lambda_2(P) \geqq \frac{2\varepsilon^{n-1}(1 - \varepsilon^n)}{3n^2(2n + 1)(1 - \varepsilon)} .$$

For a fixed $n > 2$ (i.e., a fixed problem) it is clear that the bound in (4.14) is superior to the Sinclair–Jerrum lower bound in (4.15) as $\varepsilon \rightarrow 0$. Comparison with the upper bound of (4.15) shows that the bound in (4.14) is tight when treated as a function of ε alone. On the other hand, the Diaconis–Stroock bound in (4.16) has the same exponent of ε as the bound in (4.14) but is worse as a function of n.

Remark. The methods in [22] and [10] can lead to excellent bounds in certain cases. For example, consider an SA Markov chain whose state space is the set of all matchings in a graph on $2n$ vertices, which has been analyzed in [22]. The time constant for such a chain has been shown to be $\tau \leqq f(n)\varepsilon^{-\eta}$, where $f(n)$ is independent of ε and η is constant independent of n and ε. On the other hand, the bound from Theorem 3.1 for the same chain yields $\tau \leqq f_1(n)\varepsilon^{-n}$, where $f_1(n)$ is independent of ε; this bound is clearly much worse. Thus Theorem 3.1 may be too pessimistic in some cases. Also, as remarked on earlier, Theorem 3.1 will not be directly useful when $\mu_2(Q)$ is not known.

The purpose of Example 4.1 was merely to illustrate an example of a reversible Markov chain for which the eigenvalue bound of Theorem 3.1 (hence a bound on the time constant) is tight as a function of temperature. The corresponding optimization problem, however, is very easy since, by construction, the states 1, $2n$, $2n + 1$, and $4n$ have the globally minimum cost of 1. The following example illustrates a difficult and more realistic optimization problem for which one can still use Theorem 3.1 to obtain a meaningful bound for the time constant of the corresponding SA Markov chain.

Example 4.2. Let $\{a_1 \leqq a_2 \leqq \cdots \leqq a_n\}$ be a given set of n positive integers in ascending order, and define $A = (a_1 + a_2 + \cdots + a_n)/2$. Let Ω denote the state space of all binary vectors of length n, and consider a state $u = (u_1, u_2, \ldots, u_n)$, where $u_i \in \{0, 1\}$. Define the cost of the state u as $C(u) = |A - \sum_{i=1}^{n} a_i u_i|$. The problem of finding a state of minimum cost is the optimization version of the SUBSET_SUM problem that is known to be NP-complete [11]. However, the SUBSET_SUM problem is pseudo-polynomial and can be solved by using a dynamic programming approach in time polynomial in A [11]. We will use our bounds to study the temperature asymptotics of the SA Markov chain to solve SUBSET_SUM. Define the neighbors of a state u as all states differing from u in one bit. Then the SA Markov chain has $N = 2^n$, $\rho = n + 1$, and $\Delta \leqq A$, and the underlying graph is the n-dimensional hypercube with $\mu_2(Q) = 2$. Using (4.6), we can bound the spectral gap and the time constant as

$$(4.17) \qquad 1 - \beta \geqq \frac{2\varepsilon^A}{n + 1} \quad \text{and} \quad \tau \leqq \frac{n + 1}{2} \varepsilon^{-A}$$

for any $0 < \varepsilon \leqq 1$. In computing both the Diaconis–Stroock and Sinclair–Jerrum bounds in this case, a judicious choice of canonical paths must be made. Since we are not aware of the best possible choice, we let Γ be the set of canonical paths used in [22] to compute λ_2 for a random walk on the n-dimensional hypercube. In this case each edge in the hypercube is used by 2^{n-1} paths in Γ. Computation of $U(\Gamma)$ in (4.12) requires the evaluation of a summation comprising 2^{n-1} terms for each edge in the hypercube. It is easier to see that $U(\Gamma) \leqq \rho\varepsilon^{-\Delta}\pi_{\max}2^{n-1}$. Exact calculation of $\pi_{\max}$ involves computation of the partition function; however, we will use $\pi_{\max} \leqq \min(1, 2^{-n}\varepsilon^{-\Delta})$ instead. Thus the bound on spectral gap provided by the Sinclair–Jerrum bound is

$$(4.18) \qquad 1-\beta \geqq \frac{\max(2^{-2n}, \varepsilon^{2A})}{2(n+1)^2}\varepsilon^{2A},$$

which is much worse than (4.17) both as a function of ε and as a function of n. The Diaconis–Stroock bound (using $l_{\max} = n$) gives

$$(4.19) \qquad 1-\beta \geqq \frac{2\max(2^{-n}, \varepsilon^{A})}{n(n+1)}\varepsilon^{A},$$

which is comparable to (4.17) as a function of ε (when ε is close to 0) but is worse as a function of n. Possible improvements may be obtained by a considerably more careful estimation of $U(\Gamma)$ or a different choice of Γ or both. Such efforts have provided significant rewards in the form of tight bounds (see the remark after Example 4.1) in many instances. On the other hand, when one is faced with an arbitrarily specified SA chain, Theorem 3.1 yields a useful bound in cases for which the underlying structure is well understood.

Using (1.3), (1.4), and (4.17), we see that the time k required for the SA chain of Example 4.2 to be at distance within δ from equilibrium is at most

$$(4.20) \qquad \left(\log\frac{1}{\delta} + \frac{A}{2}\log\frac{1}{\varepsilon} + \frac{n}{2}\log 2\right)\frac{n+1}{2}\varepsilon^{-A}.$$

For a fixed temperature ε the expression in (4.20) is exponential in A, which does not recommend the use of SA for solving the SUBSET_SUM problem with provably good results. This is especially the case here because there exist dynamic programming techniques that can solve SUBSET_SUM in time polynomial in A [11]. In practice, however, the SA method might perform much better than what the bounds presented in this paper allow us to show.

Acknowledgments. The authors gratefully acknowledge the suggestions, comments, and contributions of all three referees of this paper. In particular, we thank James Fill of Johns Hopkins University for refereeing this paper and for providing us with Theorem 3.5 and the exact eigenvalues in Example 3.7, along with several other invaluable comments and suggestions.

REFERENCES

[1] D. ALDOUS, *On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing*, Probab. Engrg. Inform. Sci., 1 (1987), pp. 33–46.

[2] N. ALON, *Eigenvalues and expanders*, Combinatorica, 6 (1986), pp. 83–96.

[3] N. ALON AND V. D. MILMAN, *λ_1, isoperimetric inequalities for graphs and superconcentrators*, J. Combin. Theory Ser. B, 38 (1985), pp. 73–88.

[4] L. BABAI, *Spectra of Cayley graphs*, J. Combin. Theory Ser. B, 27 (1979), pp. 180–189.

[5] N. BIGGS, *Algebraic Graph Theory*, Cambridge University Press, Cambridge, England, 1974.

[6] J. BONDY AND U. S. R. MURTY, *Graph Theory with Applications*, North-Holland, New York, 1976.

[7] J. CHEEGER, *A lower bound for the smallest eigenvalue of the Laplacian*, in Problems in Analysis, R. C. Gunning, ed., Princeton University Press, Princeton, NJ, 1970, pp. 195–199.

[8] D. P. CONNORS AND P. R. KUMAR, *Balance of recurrence order in time-inhomogeneous Markov chains with application to simulated annealing*, Probab. Engrg. Inform. Sci., 2 (1988), pp. 157–184.

[9] D. M. CVETKOVIC, M. DOOB, AND H. SACHS, *Spectra of Graphs*, Academic Press, New York, 1980.

[10] P. DIACONIS AND D. STROOCK, *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab., 1 (1991), pp. 36–61.

[11] M. R. GAREY AND D. S. JOHNSON, *Computers and Intractability; A Guide to the Theory of* NP-*Completeness*, W. H. Freeman, New York, 1979.

[12] B. HAJEK, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311–329.

[13] R. HOLLEY AND D. STROOCK, *Simulated annealing via Sobolev inequalities*, Comm. Math. Phys., 115 (1988), pp. 553–569.

[14] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.

[15] S. KIRKPATRICK, C. D. GELATT, JR., AND M. P. VECCHI, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.

[16] G. LAWLER AND A. SOKAL, *Bounds on the L^2 spectrum for Markov chains and Markov processes: A generalization of Cheeger's inequality*, Trans. Amer. Math. Soc., 309 (1988), pp. 557–580.

[17] M. P. DESAI AND V. B. RAO, *Integer compositions using simulated annealing*, Tech. Report UILU-ENG-91-2237 (DAC 28), Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 1991.

[18] D. MITRA, F. ROMEO, AND A. L. SANGIOVANNI-VINCENTELLI, *Convergence and finite-time behavior of simulated annealing*, Adv. Appl. Probab., 18 (1986), pp. 747–771.

[19] B. MOHAR, *The Laplacian spectrum of graphs*, presented at International Graph Theory Conference, Western Michigan University, Kalamazoo, MI, 1988.

[20] G. SASAKI AND B. HAJEK, *The time-complexity of maximum matching by simulated annealing*, J. Assoc. Comput. Mach., 35 (1988), pp. 387–403.

[21] E. SENETA, *Nonnegative Matrices and Markov Chains*, Springer-Verlag, Berlin, 1980.

[22] A. SINCLAIR AND M. JERRUM, *Approximate counting, uniform generation and rapidly mixing Markov chains*, Inform. and Comput., 82 (1989), pp. 93–133.

[23] M. P. DESAI AND V. B. RAO, *A Characterization of the Smallest Eigenvalue of a Graph*, Tech. Report UILU-ENG-91-2238 (DAC 29), Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, 1991.

AN ALGORITHM FOR THE LINEAR COMPLEMENTARITY PROBLEM WITH A P_0-MATRIX*

V. VENKATESWARAN†

Abstract. This paper presents an algorithm for solving the linear complementarity problem (LCP) with a P_0-matrix (i.e., a matrix with all principal minors nonnegative). The method is based on solving a perturbed problem for an arbitrarily small perturbation. To that end, a Bard-type algorithm for performing computations with rational functions is developed. The algorithm has been tested on random P_0-matrix problems that are readily constructed. The results appear to bear out the viability of the method.

Key words. LCP algorithm, P_0-matrix

AMS subject classification. 90C33

1. Introduction.

In this paper we consider the linear complementarity problem (LCP) (q, M) with a P_0-matrix:

$$\begin{gathered}\text{Find } w, z \in R^n \text{ such that}\\ w = Mz + q,\\ w, z \geq 0, \qquad w^t z = 0,\\ (M \in R^{n \times n} \cap P_0,\ q \in R^n \text{ given}).\end{gathered} \tag{1}$$

Recall that a square matrix is in the class P_0 if all of its principal minors are nonnegative. A given instance of an LCP (q, M) may or may not have a solution. By "solving an LCP" we mean either producing a (w, z) that satisfies (1) or demonstrating that no such pair of vectors exists.

Highly effective algorithms exist for important subclasses of the problem, such as when M is a P-matrix (definitions follow) or a positive semidefinite matrix [1]–[4]. However, no such algorithm exists for the general problem (1). Part of the difficulty lies in the fact that when M is in P_0, (q, M) may not have a solution even when the feasible set $F(q, M) = \{(w, z) \mid w = Mz + q; w \geq 0, z \geq 0\}$ is nonempty. The proposed algorithm does not rely on the feasible set being empty to show that the LCP has no solution. In the study of LCPs a matrix is said to belong to the class Q_0 if (q, M) has a solution whenever $F(q, M)$ is nonempty. In this terminology existing approaches will solve the LCP only when M is in Q_0.

Our method is motivated by the fact (see Theorem 1) that a simple perturbation of M results in a P-matrix. The object is to solve the perturbed problem for an arbitrarily small perturbation. To that end we develop a Bard-type algorithm [3] for working with rational functions. From the terminating tableau one can readily retrieve the solution or deduce that none exists.

Such a perturbation of the LCP when M is a P_0-matrix has been extensively studied in [5]. We comment on the relationship to that work in § 3. Also, our scheme can be thought of as a special type of regularization, and as such it is closely connected to the well-known proximal-point algorithm from monotone operator theory [6].

The Bard-type algorithm introduced by Murty in [3] has been widely studied; see, for instance, the recent papers [7] and [8]. For algorithms that use similar least-index arguments, see [9]–[11].

* Received by the editors November 28, 1990; accepted for publication (in revised form) January 4, 1992.

† AT&T Bell Laboratories, Holmdel, New Jersey (ven@hocus.att.com).

In the remainder of this section we introduce some definitions and explain our notations. In § 2 we derive the required results, and in § 3 we present the algorithm. In any practical implementation it is essential that the degree of all of the polynomials in the tableau be bounded. We give an update scheme that ensures that the degree of no polynomial exceeds n, the dimension of the problem. Finally, § 4 describes our implementation and the computational tests that we have performed with the algorithm.

Algorithms for algebraic computation are well known [12]–[14]. For an algorithm for solving a linear program when the entries in A, b, and c are polynomials, see [15].

Notation and definitions. If $M \in R^{n \times n}$ and $q \in R^n$, by $M_i(q_i)$ we mean the ith row of M (the ith coordinate of q). We use e to denote the vector of ones, with dimension to be inferred from the context. We denote $\{1, 2, \dots, n\}$ by Γ. Let $\alpha, \beta \subseteq \Gamma$. By M_α we mean the rows of M indexed by α, and by q_α we mean the subvector with components in α. We use $M_{\alpha\beta}$ to denote the submatrix of M whose rows and columns are indexed by α and β, respectively. We denote $\Gamma \backslash \alpha$ by $\bar{\alpha}$. We use $\alpha - \beta$ to denote the elements in α that are not in β. We use I to denote the identity matrix, with dimension suitable to the context. We use det (M) to denote the determinant of M. To call attention to the fact that a function f is a polynomial in ε we will write $f(\varepsilon)$. At other times, to simplify the notation, we will write f instead of $f(\varepsilon)$.

M is a *P-matrix* if all of its principal minors are positive, and it is a *positive semidefinite* (PSD) matrix if for all $x \in R^n$, $x^t M x \geq 0$. M is *strictly* (*row*) *diagonally dominant* if for every i, $M_{ii} > \sum_{j \neq i} |M_{ij}|$. A strictly diagonally dominant matrix is also a P-matrix, and a PSD matrix is also a P_0-matrix. It is well known that an LCP with a P-matrix has a unique solution for all $q \in R^n$ [16]. A square matrix M is said to belong to the class R_0 if it has a unique solution to LCP $(0, M)$. M is said to be *column sufficient* if $x_i(Mx)_i \leq 0$, $i = 1, \dots, n$, implies that $x_i(Mx)_i = 0$, $i = 1, \dots, n$, [17]. M is said to be *row sufficient* if M^T is column sufficient. M is said to be *sufficient* if it is both column and row sufficient. Column sufficient (row sufficient) matrices are a subclass of P_0-matrices.

2. The perturbed problem.

Our starting point is the following well-known result.

THEOREM 1. *M is in P_0 if and only if $(M + \varepsilon I)$ is in P for all $\varepsilon > 0$.*

Proof. See [18].

As indicated, our approach is to solve the perturbed problem $(q, (M + \varepsilon I))$. We will assume the following standard nondegeneracy condition on the q vector. With this assumption, if (1) has a solution, it has a unique basic feasible solution and the solutions of the perturbed problems converge to this solution.

Assumption 1. Every feasible basis of $F(q, M)$ is nondegenerate.

THEOREM 2 [19]. *Let $M \in P_0$, and let Assumption* 1 *hold. Let (q, M) have a solution. Then the problem has a unique basic feasible solution.*

Proof. If (q, M) has a solution, it has a basic feasible solution. Let $(-M_{\cdot K}, I_{\cdot \bar{K}})$, $K \subseteq \Gamma$, denote such a basis. For $\varepsilon > 0$ let $M(\varepsilon) = M + \varepsilon I$. By Assumption 1

$$\begin{aligned} z_K &= -[M_{KK}]^{-1} q_K > 0, \\ w_{\bar{K}} &= q_{\bar{K}} - M_{\bar{K}K} M_{KK}^{-1} q_K > 0. \end{aligned} \tag{2}$$

For sufficiently small $\varepsilon > 0$, $M(\varepsilon)_{KK}$ is nonsingular, and since (2) holds,

$$z(K)_K = -[M(\varepsilon)_{KK}]^{-1} q_K > 0,$$

$$w(K)_{\bar{K}} = q_{\bar{K}} - M(\varepsilon)_{\bar{K}K} [M(\varepsilon)_{KK}]^{-1} q_K > 0.$$

Therefore, $z(K) = (z(K)_K, 0)$, $w(K) = (0, w(K)_{\bar{K}})$ is a nondegenerate solution to

$(q, M(\varepsilon))$. By contradiction, let $(-M_{\cdot L}, I_{\cdot \bar{L}})$, $L \subseteq \Gamma$, $L \neq K$, denote a different complementary feasible basis for (q, M). Then for sufficiently small $\varepsilon > 0$, $(w(K), z(K))$ and $(w(L), z(L))$ are two distinct nondegenerate solutions to the P-matrix LCP, $(q, M(\varepsilon))$, a contradiction. □

Next, we note that if the LCP has a solution, the solution of the perturbed problem, as a function of ε, will converge to this solution as ε tends to 0.

THEOREM 3. *Let $M \in P_0$, and let Assumption* 1 *hold. Let (w^*, z^*) solve (q, M). Let $(-M_{\cdot K}, I_{\cdot \bar{K}})$ denote the unique solution basis. Let $(w(\varepsilon), z(\varepsilon))$ denote the unique solution to $(q, M(\varepsilon))$. Then for all sufficiently small $\varepsilon > 0$, $(-M(\varepsilon)_{\cdot K}, I_{\cdot \bar{K}})$ is the solution basis of $(q, M(\varepsilon))$. Furthermore, $(w(\varepsilon), z(\varepsilon)) \to (w^*, z^*)$ as $\varepsilon \to 0$.*

Proof. The proof of this theorem is similar to proof of Theorem 2; also see [5]. □

The next result shows that in the event that (q, M) has no solution, there is at least one complementary basis that is feasible for all $(q, M(\varepsilon))$ for all $\varepsilon > 0$ sufficiently small. This result is needed to ensure that the proposed algorithm terminates in a finite number of steps.

THEOREM 4. *Let $M \in P_0$. Then there exists an index set $K \subseteq \Gamma$ such that $(-M(\varepsilon)_{\cdot K}, I_{\cdot \bar{K}})$ is a feasible basis for $(q, M(\varepsilon))$ for all sufficiently small $\varepsilon > 0$.*

Proof. Consider a fixed $\varepsilon > 0$. Let $(-M(\varepsilon)_{\cdot K}, I_{\cdot \bar{K}})$, $K \subseteq \Gamma$, denote the unique solution basis for this P-matrix LCP. Then the solution $z = (z_K, 0)$, $w = (0, w_{\bar{K}})$ is given by

$$z_K = -[M(\varepsilon)_{KK}]^{-1} q_K,$$

$$w_{\bar{K}} = q_{\bar{K}} - M(\varepsilon)_{\bar{K}K}[M(\varepsilon)_{KK}]^{-1} q_K.$$

Viewed as a function of ε, each component of z_K, $w_{\bar{K}}$ is a rational function of ε. This follows from the fact that each entry of $[M(\varepsilon)_{KK}]^{-1}$ is a ratio of polynomials; the denominator is the determinant, $\det(M(\varepsilon)_{KK})$, and the numerator is a cofactor of $M(\varepsilon)_{KK}$. Since a rational function can change sign only finitely many times, as ε becomes arbitrarily small either $(-M(\varepsilon)_{\cdot K}, I_{\cdot \bar{K}})$ will remain feasible or for all $0 < \varepsilon < \bar{\varepsilon}$ and for some $\bar{\varepsilon} > 0$ it will remain infeasible. Since there are only finitely many choices for K and since for every $\varepsilon > 0$, $(q, M(\varepsilon))$ must have a solution, the result follows. □

THEOREM 5. *Let $M \in P_0$. For $\varepsilon > 0$ let $(w(\varepsilon), z(\varepsilon))$ denote the unique solution to $(q, M(\varepsilon))$. Further, let (q, M) have no solution. Then $\|w(\varepsilon^i), z(\varepsilon^i)\| \to \infty$ as $\{\varepsilon^i\} \to 0$.*

Proof. As in Theorem 4, let K denote the index set that is the unique feasible basis for $(q, M(\varepsilon))$ for all arbitrarily small $\varepsilon > 0$. By contradiction, if $\|w(\varepsilon^i), z(\varepsilon^i)\|$ is bounded, the solutions to the perturbed problems have a limit (w^*, z^*) on a subsequence. On this subsequence

$$w^*_{\bar{K}} = \lim_{\varepsilon^i \to 0} w(\varepsilon^i)_{\bar{K}} \geq 0 \qquad (\text{because } w(\varepsilon^i)_{\bar{K}} \geq 0),$$

$$z^*_K = \lim_{\varepsilon^i \to 0} z(\varepsilon^i)_K \geq 0 \qquad (\text{because } z(\varepsilon^i)_K \geq 0),$$

$$w^*_K = 0,$$

$$z^*_{\bar{K}} = 0.$$

Since $w(\varepsilon^i) = Mz(\varepsilon^i) + q$, $w^* = Mz^* + q$. Hence (w^*, z^*) solves (q, M), contradicting the hypothesis. □

In view of Theorems 3 and 4, there exists an index set $K \subseteq \Gamma$ and an $\bar{\varepsilon} > 0$ such that both of the following hold:

(i) For all $0 < \varepsilon < \bar{\varepsilon}$, $(-M(\varepsilon)_{\cdot K}, I_{\cdot \bar{K}})$ is a solution basis for $(q, M(\varepsilon))$.

(ii) Either $(-M_{\cdot K}, I_{\cdot \bar{K}})$ is a solution basis for (q, M) or the LCP has no solution. In the algorithm below we solve $(q, M(\varepsilon))$ algebraically with a view to determining this index set K.

3. Bard-type algorithm for polynomial entries. For a description of the original Bard-type algorithm for P-matrix LCPs, see [3]. Here, we develop a variant that will solve LCPs with P_0-matrices. All the entries in any intermediate tableau are taken to be rational functions. In the input data we provide $(M + \varepsilon I)$ for M; in what follows we use M to denote this perturbed matrix. The algorithm then is as follows:

Step 0. Set $B = I$. Proceed to Step 1.
Step 1. Let $\hat{q} = B^{-1}q$. Each $\hat{q}_i = P_i(\varepsilon)/Q_i(\varepsilon)$ is a rational function. Let $j = \min_{1 \le i \le n} \{i : P_i(\varepsilon)/Q_i(\varepsilon)|_{\varepsilon = 0^+} < 0\}$. If no such j exists, go to Step 3; else proceed to Step 2.
Step 2. If $-M_{\cdot j}$ is in the basis, replace it with $I_{\cdot j}$. Else, replace $I_{\cdot j}$ in the basis with $-M_{\cdot j}$. Let B denote the new basis matrix. Go to Step 1.
Step 3. Let $B = (-M_{\cdot K}, I_{\cdot \bar{K}})$, $K \subseteq \Gamma$. Let $q_i^* = \lim_{\varepsilon \to 0} \hat{q}_i$, $1 \le i \le n$. If $q_i^* = \infty$ for any $1 \le i \le n$, terminate; by Theorem 3, (q, M) has no solution. Else, $z^* = (q_K^*, 0)$, $w^* = (0, q_{\bar{K}}^*)$ is the solution by continuity, as in the proof of Theorem 5.

We make the following remarks:

1. We show below how the basis inverse and $\hat{q}$ may be updated so that the degree of no polynomial in their representation exceeds n, the dimension of the problem. Then the space required to store each rational function entry is no more than $2n + 2$, $(n + 1)$ being the space necessary to store the $(n + 1)$ coefficients of each degree-n polynomial.

2. To check whether $P_i(\varepsilon)/Q_i(\varepsilon)|_{\varepsilon = 0^+} < 0$, we proceed as follows. If a is the coefficient of the lowest-degree term in $P_i(\varepsilon)$, say, of degree k, and if b is the lowest-degree term in $Q_i(\varepsilon)$, say, of degree l, then $P_i(\varepsilon)/Q_i(\varepsilon)|_{\varepsilon = 0^+} < 0$ if and only if $a/b < 0$. Also,

$$\lim_{\varepsilon \to 0} \frac{P_i(\varepsilon)}{Q_i(\varepsilon)} = \begin{cases} 0 & \text{if } k > l, \\ \infty & \text{if } l > k, \\ a/b & \text{if } k = l; \end{cases}$$

cf. Step 3.

3. The index j chosen in Step 1 is precisely the least index prescribed in [3]. As in the proof of Theorem 4, each $\hat{q}_i$ is a rational function of ε and can change sign only finitely many times. Thus for each basis B there exists an $\bar{\varepsilon}(B)$ such that, for all $0 < \varepsilon < \bar{\varepsilon}(B)$, the index j (Step 1) is the least index of [3]. Since there are only a finite number of bases, there exists an $\bar{\varepsilon} > 0$ such that, for all $0 < \varepsilon < \bar{\varepsilon}$, Step 1 produces the least index for every basis.

The algorithm in [3] terminates in a finite number of steps when the matrix is a P-matrix. In view of the fact that $(M + \varepsilon I) \in P$ for all $\varepsilon > 0$ and of the remarks following Theorem 5, the above algorithm also terminates in a finite number of steps.

To implement the algorithm it is sufficient to keep B^{-1} and $\hat{q}$. Entries of B^{-1}, $\hat{q}$ are rational functions. We will provide an update scheme that ensures that the denominators of all of the entries in B^{-1}, $\hat{q}$ are identical and, in fact, are equal to det (B). With this invariant it is easy to see that the numerators of the entries in B^{-1}, $\hat{q}$ are polynomials of degree no greater than n. If it is assumed that this invariant holds at some stage, Theorem

6 shows how it can be maintained during a pivot. The key is to perform certain divisions. We show, moreover, that in each case the divisor divides the dividend exactly, without remainder. (Note that in terms of data structures for storing B^{-1}, $\hat{q}$, it suffices to keep only the numerators since the denominators are all identical.)

THEOREM 6. *Let* $B = (-M_{\cdot K}, I_{\cdot \bar{K}})$, $K \subseteq \Gamma$, *denote a complementary basis. Let* $B^{-1} = [b_{kl}]$. *For* $1 \le k,\ l \le n$, *in the representation of* b_{kl} *as a ratio of polynomials let the denominator equal* $d(\varepsilon) = \det(B)$. *Let* $i \in \Gamma$. *If* $i \in K$, *let* $u = I_{\cdot i}$ *and* $K' = K - \{i\}$. *If* $i \in \bar{K}$, *let* $u = -M_{\cdot i}$ *and* $K' = K \cup \{i\}$. *Let* $B' = (-M_{\cdot K'}, I_{\cdot \bar{K}'})$. *Let*

$$b_{kj} := \frac{p_{kj}(\varepsilon)}{d(\varepsilon)}, \qquad 1 \le k, j \le n,$$

$$\hat{u}_k = (B^{-1}u)_k := \frac{v_k(\varepsilon)}{d(\varepsilon)}, \qquad 1 \le k \le n,$$

$$\hat{q}_k = (B^{-1}q)_k := \frac{r_k(\varepsilon)}{d(\varepsilon)}, \qquad 1 \le k \le n.$$

Then the following hold:

(i) $v_i = \det(B')$,

(ii) $d \mid (p_{kj}v_i - p_{ij}v_k)$ *and* $d \mid (r_k v_i - r_i v_k)$.

Proof. (i) *Case* 1 ($i \in K$). Since

$$u = I_{\cdot i} \quad \text{and} \quad B^{-1} = \begin{bmatrix} -M_{KK}^{-1} & 0 \\ -M_{\bar{K}K}M_{KK}^{-1} & I_{\bar{K}\bar{K}} \end{bmatrix},$$

$$\begin{aligned} \hat{u}_i &= [B^{-1}]_{ii} \\ &= \frac{(B)_{ii}}{\det(B)} \\ &= \frac{\det(B')}{\det(B)}, \end{aligned}$$

where $(B)_{ii}$ is the (i, i) cofactor of B. Therefore, $\hat{u}_i = v_i/d$ implies that $v_i = \det(B')$.

Case 2 ($i \in \bar{K}$). Then $u = -M_{\cdot i}$, and

$$\begin{aligned} \hat{u}_i &= -[B^{-1}M_{\cdot i}]_i, \\ &= \frac{\det \begin{bmatrix} -M_{K'K'} & 0 \\ -M_{\bar{K}'K'} & I_{K'K'} \end{bmatrix}}{\det(B)} \qquad \text{(by Cramer's rule)}, \\ &= \frac{\det(B')}{\det(B)}. \end{aligned}$$

Again, $\hat{u}_i = v_i/d$ implies that $v_i = \det(B')$.

(ii) Let $[B']^{-1} = [b'_{kj}]$. Then, for $1 \le k, j \le n$, each entry b'_{kj} can be expressed as a rational function, with denominator $d' = \det(B')$ and numerator an appropriate cofactor of B', p'_{kj}. Hence if $b'_{kj} = p'_{kj}/d'$, by pivoting on $\hat{u}_i$, b'_{kj} can also be expressed as

$$\frac{p_{kj}}{d} - \frac{v_k}{d}\frac{p_{ij}}{d}\frac{d}{d'} = \frac{p_{kj}d' - v_k p_{ij}}{dd'} \qquad (\text{since } v_i = d').$$

Therefore, $d \mid (p_{kj}v_i - v_k p_{ij})$.

Similarly, if $\hat{q}'_k = ([B']^{-1}q)_k = r'_k/d'$, again by pivoting, $\hat{q}'_k$ can be expressed as

$$\frac{r_k}{d} - \frac{v_k}{d}\frac{r_i}{d}\frac{d}{d'} = \frac{r_k d' - r_i v_k}{dd'}.$$

Therefore, $d|(r_k v_i - r_i v_k)$. □

In view of the above theorem, with $v = [p_{ij}]u$, $[p_{ij}]$, r, and d can be updated as follows:

$$p_{kj} \leftarrow \frac{p_{kj}v_i - p_{ij}v_k}{d}, \qquad 1 \le k, j \le n, \quad k \ne i,$$

$$r_k \leftarrow \frac{r_k v_i - r_i v_k}{d}, \qquad 1 \le k \le n, \quad k \ne i,$$

$$d \leftarrow v_i.$$

We make the following observations:

1. With $B = I$, evidently the property in the hypothesis of Theorem 6 holds. In all subsequent iterations the property holds by Theorem 6 and induction. With this scheme, since each p_{ij} is an appropriate cofactor of B, its degree does not exceed n. Since $r_i = \sum_{j=1}^{n} p_{ij}q_j$, its degree is also bounded by n. Clearly, the degree of $d = \det(B)$ also does not exceed n.

2. It is easy to see that d above is always a monic polynomial. (This follows from our perturbation scheme $M \leftarrow M + \varepsilon I$.) Hence in evaluating p_{kj}, r_k above, no numeric division is involved. The algorithm is, therefore, an all-integer, exact algorithm.

3. In the terminal iteration, if the polynomial d has no constant term, at $\varepsilon = 0$ the set of columns representing the basis is no longer linearly independent. By Theorem 3 (and Assumption 1), (q, M) has no solution. Conversely, if $d(\varepsilon)$ has a constant term, $d(0) \ne 0$ and the basis is also a basis of (1). Also, the entries in q^* are then finite and clearly $w = (0, q_{\bar{K}}^*)$, $z = (q_K^*, 0)$ solves (1) by continuity.

We illustrate the algorithm with a numerical example.

Example.

$$M = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \qquad q = \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix}.$$

Iteration 1.

$$B = I, \quad d = 1, \quad [p_{ij}] = I, \quad r = \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix},$$

$$\hat{q} = \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix}.$$

The entering variable is z_1.

Iteration 2.

$$B=\begin{bmatrix}-1-\varepsilon & 0 & 0\\ 1 & 1 & 0\\ 0 & 0 & 1\end{bmatrix},\quad v=\begin{bmatrix}-1-\varepsilon\\ 1\\ 0\end{bmatrix},\quad d=-1-\varepsilon,$$

$$[p_{ij}]=\begin{bmatrix}1 & 0 & 0\\ -1 & -1-\varepsilon & 0\\ 0 & 0 & -1-\varepsilon\end{bmatrix},\quad r=\begin{bmatrix}-2\\ 3+\varepsilon\\ -1-\varepsilon\end{bmatrix},$$

$$\hat{q}=\begin{bmatrix}\dfrac{-2}{-1-\varepsilon}\\ \dfrac{3+\varepsilon}{-1-\varepsilon}\\ 1\end{bmatrix}.$$

The entering variable is z_2.

Iteration 3.

$$B=\begin{bmatrix}-1-\varepsilon & -1 & 0\\ 1 & -\varepsilon & 0\\ 0 & 0 & 1\end{bmatrix},\quad v=\begin{bmatrix}-1\\ 1+\varepsilon+\varepsilon^2\\ 0\end{bmatrix},\quad d=1+\varepsilon+\varepsilon^2,$$

$$[p_{ij}]=\begin{bmatrix}-\varepsilon & 1 & 0\\ -1 & -1-\varepsilon & 0\\ 0 & 0 & 1+\varepsilon+\varepsilon^2\end{bmatrix},\quad r=\begin{bmatrix}-1+2\varepsilon\\ 3+\varepsilon\\ 1+\varepsilon+\varepsilon^2\end{bmatrix},$$

$$\hat{q}=\begin{bmatrix}\dfrac{-1+2\varepsilon}{1+\varepsilon+\varepsilon^2}\\ \dfrac{3+\varepsilon}{1+\varepsilon+\varepsilon^2}\\ 1\end{bmatrix}.$$

The entering variable is w_1.

Iteration 4.

$$B=\begin{bmatrix}1 & -1 & 0\\ 0 & -\varepsilon & 0\\ 0 & 0 & 1\end{bmatrix},\quad v=\begin{bmatrix}-\varepsilon\\ -1\\ 0\end{bmatrix},\quad d=-\varepsilon,\ [p_{ij}]=\begin{bmatrix}-\varepsilon & 1 & 0\\ 0 & 1 & 0\\ 0 & 0 & -\varepsilon\end{bmatrix},\quad r=\begin{bmatrix}-1+2\varepsilon\\ -1\\ -\varepsilon\end{bmatrix},$$

$$\therefore \hat{q}=\begin{bmatrix}\dfrac{-1+2\varepsilon}{-\varepsilon}\\ \dfrac{1}{\varepsilon}\\ 1\end{bmatrix}.$$

The terminating criterion is satisfied. Since $d(0)=0$, the problem has no solution. We note that for this problem, $F(q, M)$ is in fact nonempty.

We now comment on the assumption that we have used. To conclude that LCP (q, M) has no solution when the solutions to LCP $(q, M(\varepsilon))$ diverge, we need to impose further conditions on the problem. Assumption 1 is one possible condition. Some others are the following:

(i) Let $B = \lim_{\varepsilon \to 0^+} B(\varepsilon)$, where $B(\varepsilon)$ is the solution basis to LCP $(q, M(\varepsilon))$. Then we require $\{x \in R^n | Bx = 0, x \geq 0\} = \{0\}$. This is the condition derived by Gana in [5].

(ii) $M \in R_0$.

(iii) $M \in$ PSD, and LCP (q, M) is solvable.

Note that whereas all of the above are sufficient conditions, none are necessary. Indeed, our example above does not satisfy any of the conditions (i)–(iii). However, it satisfies Assumption 1, and therefore we are able to conclude that the LCP does not have a solution when the solutions to the perturbed problems diverge.

In our development we have used Assumption 1 since it is simple and has the property that given a matrix $M \in P_0$, it is satisfied for almost all $q \in R^n$.

4. Algorithm implementation. Our code is written in the C programming language and makes use of doubly linked lists to store and manipulate polynomials. As indicated earlier, the algorithm is an all-integer, exact algorithm. The code is designed for the IBM 3090 computer; we perform all arithmetic operations with a radix of 10^{16}. Polynomial multiplication is done by first computing the convolution of the coefficients. To carry out the multiple-precision arithmetic operations involved we have used the classical algorithms from [20]. The only division operation that is performed is in computing (z^*, w^*) finally (Step 3). For this we use an approximate multiple-precision division algorithm.

It is not straightforward to generate a P_0-matrix that is also neither a P-matrix nor a positive semidefinite matrix. Our method has been to generate a lower-triangular matrix with a nonnegative (but not positive) diagonal. By making the off-diagonal entries suitably large, we prevent the matrix from becoming positive semidefinite. Also, as mentioned earlier, sufficient and column-sufficient matrices are contained in the class P_0. We have used a method proposed in [11] to construct simple test matrices of these types as well. In addition, to check the robustness of our procedure we also tested the algorithm on certain P and positive semidefinite matrices. To exercise the algorithm thoroughly some of these latter problems were constructed not to have solutions. We give below complete details of how these problems were generated. In each set, we tested the algorithm on problems of size $n = 10$, 20, and 30. Entries of M and q were all integers with three significant digits. Average computation times are shown in Table 1. In each case the time reported is the average for 10 problems. Solutions from the first two sets were verified by using Lemke's algorithm implemented as in [21]. Also, it is known that when M is a sufficient matrix, the LCP can be solved by Lemke's algorithm [17]. This provided a check of our solution for those instances. For the column-sufficient matrices (designed not to be row sufficient) we had no adequate method of verifying our solution. However, we did run Lemke's algorithm on these problems as well. In no instance did it find a solution when our algorithm had failed to find one.

1. *Strictly diagonally dominant matrices.* Off-diagonal entries were obtained by sampling from the uniform distribution $U[-1000, 1000]$ and truncating to integers. Diagonal entries were determined as

$$M_{ii} = \sum_{\substack{j=1 \\ j \neq i}}^{n} |M_{ij}| + U[0, 50].$$

Entries for q were drawn from $U[-1000, 1000]$.

TABLE 1
Average computation times for different problem types and dimensions.

	CPU seconds		
	$n = 10$	$n = 20$	$n = 30$
1 Strictly diagonal dominant M	0.94	41.53	420.13
2(a) PSD M	0.91	58.6	361.67
2(b) PSD M without solution	3.68	92.93	643.23
3(a) Special P_0	0.33	7.45	42.45
3(b) Special P_0 without solution	1.27	21.83	121.51
4(a) Sufficient M	8.36	27.76	224.97
4(b) Column-sufficient M	1.45	22.09	156.33

2. *Positive semidefinite matrices*. The first set of problems that we tried involved random symmetric positive semidefinite matrices and random q vectors. Some of the problem instances had solutions, whereas others did not. To further exercise the algorithm we put together a second set of problems without solutions. Results for these two sets of problems are shown under the headings "PSD" and "PSD without solution" in Table 1. Details on how these problems were generated follow.

2(a). PSD. M was constructed to be symmetric positive semidefinite by setting

$$M = AA^T \qquad (A : \text{lower triangular, PSD}).$$

Roughly a third of the diagonal entries of A were randomly set to zero. Off-diagonal entries of A were generated as

$$A_{ij} \sim U[-d, d] \qquad (d = 3000.0/n,\ n = \text{dimension of the problem}).$$

The limit d, above, is designed to keep the largest entry in M within $[-1000, 1000]$. The expression for d follows from the fact the variance of $U[a, b]$ is $(b-a)^2/12$. Entries for q were drawn from $U[-1000, 1000]$.

2(b). PSD *without solution*. First the matrix M was constructed as above. Then a vector d satisfying

$$M^T d \leq 0, \quad d \geq 0, \quad e^t d = 1$$

was obtained (if one existed) by solving an LP. Then q was constructed as

$$q_i = \begin{cases} -d_i & \text{if } d_i > 0, \\ \sim U[-1000, 1000] & \text{if } d_i = 0. \end{cases}$$

With this construction $q^t d < 0$ and the LCP has no solution.

3. *Special P_0-matrices*. M was constructed to be lower triangular with a nonnegative diagonal. Roughly a third of the diagonal entries were randomly set to zero. Off-diagonal entries were chosen from $U[500, 1000]$ or $U[-1000, -500]$ with a probability of 0.5. Nonzero diagonal entries were drawn from $U[1, 300]$. A matrix M constructed in this manner is in P_0 but, in general, is not positive semidefinite. Entries for q were drawn from $U[-1000, 1000]$. Obviously, such LCPs can be solved by backsubstitution.

Step 0. $\hat{q} \leftarrow q$, $i \leftarrow 1$.

Step 1. If $\hat{q}_i < 0$, $M_{ii} = 0$, LCP has no solution; terminate. Else, set $z_i = -\hat{q}_i / M_{ii}$, $w_i = 0$. If $\hat{q}_i \geq 0$, set $z_i = 0$, $w_i = \hat{q}_i$. Set $\hat{q} \leftarrow \hat{q} + M_{\cdot i} z_i$. Set $i \leftarrow i + 1$. If $i \leq n$, repeat Step 1. Else, terminate with a solution (w, z).

It is clear that if $\hat{q}_i = 0$ in Step 1 and if q is also in $F(q, M)$, it is degenerate. Hence to satisfy Assumption 1, if $\hat{q}_i = 0$ in Step 1, we perturb q_i by setting $q_i \leftarrow q_i + 1$. Results for problems with and without solutions are reported separately in Table 1 under the headings "Special P_0" and "Special P_0 without solution." In every instance, the algorithm either produced a solution or correctly showed that none exists. The Lemke algorithm terminated on a ray in each case. For the problems with a solution, this confirms that the matrix was not positive semidefinite. Among the problems with no solution, there were eight with nonempty feasible sets (three each with $n = 10$, 20 and two with $n = 30$).

4. *Sufficient and column-sufficient matrices*. In [11] it is shown that if N_{11} is sufficient (column sufficient), the matrix M constructed as follows is also sufficient (column sufficient):

$$M = \begin{bmatrix} N_{11} & \cdots & N_{1n} & a \\ \cdot & & \cdot & 0 \\ \vdots & & \vdots & \vdots \\ N_{n1} & \cdots & N_{nn} & 0 \\ b\ 0 & \cdots & 0 & c \end{bmatrix} \qquad a \cdot b < 0, \quad c \geq 0.$$

To construct a sufficient matrix we start with a 1 × 1 sufficient matrix $[N_{11}]$ and inductively build it up to size $n \times n$ by using the above result. At each stage, $|a|$, $|b|$, and c were drawn from $U[1, 1000]$ and c was set to 0 with a probability of 0.25. (The initial seed N_{11} was also chosen likewise.) To generate a column-sufficient matrix, we start instead with a 2 × 2 matrix N_{11} of the form

$$\begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix}, \qquad a \sim U[1, 1000], \quad b \sim U[-1000, 1000].$$

Note that such a matrix is column sufficient but not sufficient. In the case of the sufficient-matrix problems, entries for q were randomly drawn from $U[-1000, 1000]$. In the case of the column sufficient matrix problems, on the other hand, q was chosen so that the problems would have solutions, with high probability. With M constructed as above, strictly complementary solution vectors $z_i^* \sim U[0, 1]$, $w_i^* \sim U[1, 1000]$ are generated and q is set equal to $(w - Mz)$. (Each z_i^* is nonzero with probability 0.5.) However, since the input to our algorithm must be all integer (obtained from M, q by rounding down), the procedure does not always guarantee a solution.

The results for the sufficient-matrix problems were verified in each case by using Lemke's algorithm. As constructed, these problems tend not to have solutions, i.e., 60% of the problems with $n = 10$ did not have solutions, and none of the problems had solutions for $n = 30$.

With the column-sufficient-matrix problems, our algorithm found solutions in 24 instances (out of 30). As mentioned earlier, our procedure is not guaranteed to produce problems with solutions. Also, since these matrices have a zero column and since there is a 0.5 probability of using this column in generating q, these problems tend not to satisfy Assumption 1. In such instances our algorithm is not guaranteed to find a solution.

As mentioned earlier, we also ran Lemke's algorithm on each problem. It produced a solution in only 11 instances. In no instance did Lemke's algorithm produce a solution when our algorithm had failed to find one.

In summary, from these computational tests it would appear that the method presented here holds promise as a means for solving LCPs with P_0-matrices.

Acknowledgments. I would like to thank Rick Stone and Robert Freund for helpful discussions. I also wish to thank the referees for many constructive comments.

REFERENCES

[1] C. E. Lemke, *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11 (1965), pp. 681–689.

[2] R. W. Cottle and G. B. Dantzig, *Complementary pivot theory of mathematical programming*, Linear Algebra Appl., 1 (1968), pp. 103–125.

[3] K. G. Murty, *Note on a Bard-type scheme for solving the complementarity problem*, Opsearch, 11 (1974), pp. 123–130.

[4] R. L. Graves, *A principal pivoting simplex algorithm for linear and quadratic programming*, Oper. Res., 15 (1967), pp. 482–494.

[5] A. Gana, *Studies in the Complementarity Problem*, Ph.D. dissertation, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI, 1982.

[6] R. W. Cottle, J-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, San Diego, CA, 1992.

[7] J. Judice and F. M. Pires, *Bard-type methods for the linear complementarity problem with symmetric positive definite matrices*, Res. Report, Departimento de Matematica, Universidade de Coimbra, Coimbra, Portugal, 1988.

[8] J. Rohn, *A short proof of finiteness of Murty's principal pivoting algorithm*, Math. Programming, 46 (1990), pp. 255–257.

[9] Y.-Y. Chang, *Least index resolution of degeneracy in linear complementarity problems*, Tech. Report 79-14, Department of Operations Research, Stanford University, Stanford, CA, 1979.

[10] E. Klafszky and T. Terlaky, *Some generalizations of the criss-cross method for the linear complementarity problem of oriented matroids*, Combinatorica, 9 (1989), pp. 189–198.

[11] R. W. Cottle and Y.-Y. Chang, *Least index resolution of degeneracy in linear complementarity problems with sufficient matrices*, SIAM J. Matrix Anal. Appl., to appear.

[12] B. Buchberger, G. E. Collins, and R. G. K. Loos, eds., *Computer Algebra—Symbolic and Algebraic Computation*, 2nd ed., Springer-Verlag, Vienna, 1983.

[13] J. H. Davenport, Y. Siret, and E. Tournier, *Computer Algebra Systems and Algorithms for Algebraic Computation*, Academic Press, New York, 1988.

[14] G. Rayna, REDUCE—*Software for Algebraic Computation*, Springer-Verlag, New York, 1987.

[15] E. Weickenmeier, *Zur Lösung parametrischer linearer Programme mit polynomischen Parameterfunktionen*, Z. Oper. Res., 22 (1978), pp. 131–149.

[16] H. Samelson, R. M. Thrall, and O. Wesler, *A partition theorem for Euclidean n-space*, Proc. Amer. Math. Soc., 9 (1958), pp. 805–807.

[17] R. W. Cottle, J.-S. Pang, and V. Venkateswaran, *Sufficient matrices and the linear complementarity problem*, Linear Algebra Appl., 114/115 (1989), pp. 231–249.

[18] M. Fiedler and V. Ptak, *Some generalizations of positive definitiveness and monotonicity*, Numer. Math., 9 (1966), pp. 163–172.

[19] B. C. Eaves, *The linear complementarity problem*, Management Sci., 17 (1971), pp. 612–634.

[20] D. E. Knuth, *The Art of Computer Programming*, Vol. 2, 2nd ed., Addison-Wesley, Reading, MA, 1981.

[21] *Collected* ACM *Algorithms*, Association for Computing Machinery, New York, 1981.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 978–990, October 1993

ROBUST CONTROL OF DELAY FEEDBACK SYSTEMS WITH BOUNDED UNCERTAINTY*

R. T. YANUSHEVSKY†

Abstract. The problem of the analytical design of controllers for a class of delay plants with uncertain parameters is considered. The proposed procedure of the synthesis of a class of robust time-invariant linear and nonlinear control systems with delay is based on the consideration of an optimal control problem with a specified performance index for a linear plant with specified parameters.

Key words. linear delay systems, nonlinear delay systems, robust delay systems, robust control

AMS subject classifications. 15A24, 15A45, 93

Introduction. Robustness of control systems refers to the ability of the system to maintain stability and/or performance characteristics in the presence of parameter uncertainties. Intensive research in the area of robust control was inspired by the Kharitonov theorem, which is related to the asymptotic stability of a family of systems described by linear differential equations [12]. Different generalizations of the Kharitonov theorem for continuous and discrete linear systems in the time and frequency domains have been discussed in [2]–[5], [7], [9], [10], [16], [17].

Based on criteria similar to the well-known criteria in the classical automatic control theory, the problem of robust stability in linear time-invariant closed-loop systems with parametric uncertainty is formulated essentially as the analysis problem. By using the Kharitonov-type theorem, we can only check if the system with perturbed parameters remains stable. There exists no effective design procedure of changing the controller parameters to make the real (not nominal) system asymptotically stable.

Parallel with the approach mentioned above, the Lyapunov–Bellman method is used to design controllers for systems with uncertain dynamics (see, e.g., [1], [8], [14], [15], [22], [23]). In [14], a measure of stability robustness was presented in terms of the solution of a Lyapunov matrix equation. A modified Riccati equation is used in [14], [15], [22], [23] to design the controller, which guarantees robust stability. The robustness problem is combined with the known H_2 and H_∞ problems.

Unlike the literature devoted to robustness of systems described by ordinary differential equations, we can point out only some papers related to the robustness problem for systems with delays [6], [11] which generalize results obtained in [4], [12].

This paper presents the procedure of the robust delay control systems design. The proposed approach differs from the techniques based on the use of the Kharitonov-type theorems. It is based on the consideration of an optimal control problem with a specified performance index.

Statement of the problem and main results. Let us consider the linear uniformly controllable plant described by the equation

$$\begin{aligned} \dot{x}(t) &= \sum_{i=0}^{l} A_i x(t-\tau_i) + Bu(t) \\ x(t) &= \phi_x(t), \quad -\tau \leqq t \leqq 0, \end{aligned} \tag{1}$$

where x is an m-dimensional state space vector; u is an n-dimensional control vector;

* Received by the editors August 6, 1990; accepted for publication (in revised form) January 6, 1992.

† Department of Mechanical Engineering, University of Maryland, College Park, Maryland 20742.

$A_i = [a_{ikj}]$ and $B = [b_{kj}]$ are matrices of appropriate dimensions; $0 = \tau_0 < \cdots < \tau_l = \tau$ are time delays associated with the system coordinates; and $\phi_x(t)$ is the initial function of $x(t)$.

It is assumed first that only elements of the state matrices are not known exactly, i.e.,

$$A_{1i} \leqq A_i \leqq A_{2i}, \tag{2}$$

where $A_{1i} = [a_{ikj}^{(1)}]$ and $A_{2i} = [a_{ikj}^{(2)}]$ characterize the upper and lower bounds of A_i, respectively (the inequalities (2) are componentwise).

The robust control problem consists in finding state feedback controller equations, which make the closed-loop system asymptotically stable for all state matrices of the form (2).

The well-known procedure of analytical controller design for systems of the form (1) is based on minimization of the functional

$$J_0 = \frac{1}{2} \int_0^\infty (x^T(t) Q x(t) + c u^T(t) u(t))\, dt, \tag{3}$$

where $Q = [q_{ij}]$ is a nonnegative definite symmetric matrix and c is a positive constant (see, e.g., [19], [20]).

The optimal control law has the form [19], [20]

$$u(t) = -\frac{1}{c} B^T \left[W x(t) + \int_{-\tau}^{0} B(\xi) x(t+\xi)\, d\xi \right], \tag{4}$$

where the matrices $B(\xi)$ and W, together with $P(\xi, \sigma)$, satisfy the Riccati equations

$$Q + A_0^T W + W A_0 - \frac{1}{c} W B B^T W + B(0) + B^T(0) = 0, \tag{5}$$

$$\frac{dB(\xi)}{d\xi} - A_0^T B(\xi) - P(0, \xi) + \frac{1}{c} W B B^T B(\xi) = 0, \quad -\tau \leqq \xi \leqq 0, \tag{6}$$

$$\frac{\partial P(\xi, \sigma)}{\partial \xi} + \frac{\partial P(\xi, \sigma)}{\partial \sigma} + \frac{1}{c} B^T(\xi) B B^T B(\sigma) = 0, \quad -\tau \leqq \xi \leqq 0, \quad -\tau \leqq \sigma \leqq 0, \tag{7}$$

$$W A_i - B(-\tau_i^+) + B(-\tau_i^-) = 0, \quad (i = 1, \ldots, l-1), \tag{8}$$

$$A_i^T B(\xi) - P(-\tau_i, \xi) + P(-\tau_i, \xi) = 0, \quad (i - 1, \ldots, l-1), \tag{9}$$

$$A_l^T B(\xi) - P(-\tau, \xi) = 0, \quad W A_l - B(-\tau) = 0. \tag{10}$$

(The above equations correspond to the particular case of the expressions given in [19]; the upper indices "+" and "−" denote the left ($\tau_i - 0$) and the right ($\tau_i + 0$) limits at the points of discontinuity.)

The minimal value of (3) $V(x(\xi))$ follows from

$$\begin{aligned} 2V(t+\xi) = {} & x^T(t) W x(t) + \int_{-\tau}^{0} (x^T(t) B(\xi) x(t+\xi) + x^T(t+\xi) B^T(\xi) x(t))\, d\xi \\ & + \int_{-\tau}^{0} \int_{-\tau}^{0} x^T(t+\xi) P(\xi, \sigma) x(t+\sigma)\, d\sigma\, d\xi. \end{aligned} \tag{11}$$

It is known [20] (see also Appendix 1) that if the control law (4) minimizes the functional (3) subject to system (1), then the control law

$$u(t) = -\frac{1}{c} B^T \left[W x(t) + \int_{-\tau}^{0} B(\xi) e^{\gamma \xi} x(t+\xi)\, d\xi \right] \tag{12}$$

minimizes the functional

$$J_1 = \frac{1}{2}\int_0^\infty e^{2\gamma t}(x^T(t)Qx(t) + cu^T(t)u(t))\,dt \tag{13}$$

subject to the system (1).

We will consider how this property can be used to solve the problem of robust control of system (1).

In practice, linear differential-difference systems are analysed by considering their approximation. Moreover, the solution of the Riccati equation (5)–(10) is obtained approximately on the basis of a discrete approximation. In the future, we will use the finite-dimensional approximation of system (1)

$$\begin{gathered}\dot{x}_0(t) = A_0x_0(t) + \sum_{i=1}^{l} A_i z_{\tau_i}(t) + Bu(t),\\ \frac{\tau_l}{N}\dot{z}_i(t) + z_i(t) = z_{i-1}(t) \quad (i = 1, \dots, N),\\ z_0(t) = x_0(t), \quad z_{\tau_i}(t) = z_{N(\tau_i/\tau_l)}(t),\end{gathered} \tag{14}$$

and the difference approximation of (5)–(10) (we do not give here the boundary and discontinuity conditions; see [19], [20])

$$Q + A_0^TW + WA_0^T - \frac{1}{c}WBB^TW + B[0] + B^T[0] = 0, \tag{15}$$

$$\frac{N}{\tau_l}(B[1-i] - B[-i]) - A_0^TB[1-i] - P[0, 1-i] + \frac{1}{c}WBB^TB[1-i] = 0, \tag{16}$$

$$\begin{aligned}&\frac{N}{\tau_l}(P[1-i, 1-j] - P[-i, 1-j] + P[1-i, 1-j] - P[1-i, -j])\\ &\qquad + \frac{1}{c}B^T[1-i]BB^TB[1-j] = 0,\end{aligned} \tag{17}$$

where $x_0(t)$ and $z(t)$ are m-dimensional vectors; N is an integer.

By introducing the state and input matrices

$$A_N = \begin{bmatrix} A_0 & \cdots & 0 & A_1 & \cdots & 0 & A_l \\ \frac{N}{\tau_l}I & \cdots & 0 & 0 & \cdots & 0 & 0 \\ \cdots & & \cdots & & & \cdots & \\ 0 & \cdots & \frac{N}{\tau_l}I & -\frac{N}{\tau_l}I & \cdots & 0 & 0 \\ \cdots & & \cdots & & & \cdots & \\ 0 & \cdots & 0 & 0 & \cdots & \frac{N}{\tau_l}I & -\frac{N}{\tau_l}I \end{bmatrix}; \quad B_N = \begin{bmatrix} B \\ 0 \\ \cdots \\ 0 \\ 0 \end{bmatrix} \tag{18}$$

and the positive definite matrix

$$W_N = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1,N+1} \\ W_{21} & W_{22} & \cdots & W_{2,N+1} \\ \cdots & \cdots & \cdots & \cdots \\ W_{N+1,1} & W_{N+1,2} & \cdots & W_{N+1,N+1} \end{bmatrix} \tag{19}$$

with elements $W_{11} = W$, $(N/\tau_l)W_{1i} = B[1-i]$, and $(N/\tau_l)^2 W_{ij} = P[1-i, 1-j]$, the approximate system (14) with the state vector $x_N(t) = \{x_0(t), z_1(t), \dots, z_N(t)\}$, the system of difference equations (15)–(17), the optimal control $u_N[x_N(t)]$ in this system with respect to the functional (3), and the approximate optimal value V_N of the performance index (11) can be presented in the form [19], [20]

$$\dot{x}_N(t) = A_N x_N(t) + B_N u_N(t), \tag{20}$$

$$u_N(t) = -\frac{1}{c} B_N W_N x_N(t), \tag{21}$$

$$Q_N + A_N^T W_N + W_N A_N - \frac{1}{c} W_N B_N B_N^T W_N = 0, \tag{22}$$

$$V_N(x_0, z_i) = \frac{1}{2} x_N^T W_N x_N, \tag{23}$$

where $u_N(t)$ denotes the control of the approximate system and

$$Q_N = \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix}. \tag{24}$$

The next consideration is based on the following theorems [19], [20].

THEOREM 1. *Let $u[x(t)]$, $x[u(t)]$, $V(x(\xi))$, and $u_N[x_N(t)]$, $x_N[u_N(t)]$, V_N be the optimal solutions of the original* (1), (4), (11) *and approximate* (20), (21), (23) *quadratic cost optimal problems. Then, for any arbitrary small $\delta > 0$, a number N_0 can be found such that for all $N \geqq N_0$*

$$\int_0^\infty \|x(t) - x_0(t)\|^2 \, dt < \delta, \tag{25}$$

$$|V(x(\xi)) - V_N| \leqq 0.5(1+q)\delta. \tag{26}$$

(*The symbol $\|\cdot\|^2$ denotes the square of the euclidean norm of the corresponding vector*; $q = \max_{ij} q_{ij}$.)

This theorem justifies the approximation of the type (14)–(23) for the study of the given class of delayed systems.

The concept of the strong dominant matrices and the dominant matrices in the following discussion is based on comparing the performance of optimal systems (the minimal values of the cost functionals) with various state matrices.

Let $W_{ON} = W_N + \Delta W_N$, $A_{*i} = A_i + \Delta A_i$, $A_{*N} = A_N + \Delta A_N$ $(i = 0, 1, \dots, l)$ and

$$\Delta A_N = \begin{bmatrix} \Delta A_0 & \cdots & 0 & \Delta A_1 & \cdots & 0 & \Delta A_l \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where W_{ON} and W_N correspond to the optimal solutions (22) that are the approximations of the optimal solutions for the original system (1) with the state matrices $A_{*0} + \gamma I$, $A_{*i}e^{\gamma\tau_i}$, and A_i, respectively; A_{*i} and A_i are state matrices of family (2); ΔA_i, ΔW_N, and ΔA_N are some matrices, with ΔA_N being formed from A_{*N} (see (18)) corresponding to A_{*i} $(i = 0, 1, \dots, l)$.

THEOREM 2. *Suppose W_{ON} and W_N correspond to the optimal solutions* (22) *for the approximate systems* (14) ((20)) *with the state submatrices $A_{*0} + \gamma I$ and $A_{*i}e^{\gamma\tau_i}$, and with the state submatrices $A_i = A_{*i} - \Delta A_i$ $(i = 1, \dots, l)$ of family* (2). *If there exist N_0 and $\gamma \geqq 0$ such that, for all $N \geqq N_0$,*

$$Q_N(W_{ON}) = 2\gamma W_{ON} + \Delta A_N^T W_{ON} + W_{ON}\Delta A_N \tag{27}$$

is a positive definite matrix for all ΔA_i *corresponding to* A_i *of family* (2), *then* $W_{ON} - W_N > 0$ *for all* A_i *of family* (2) *in the case* $\gamma > 0$ *and for all* $A_i \neq A_{*i}$ *in the case* $\gamma = 0$.

Proof. The optimal values W_{ON} and W_N of (3) corresponding to the matrices $A_{*0} + \gamma I$, $A_{*i}e^{\gamma \tau_i}$, and arbitrary matrices of family (2), respectively, satisfy the Riccati equations

$$Q_N + 2\gamma W_{ON} + A_{*N}^T W_{ON} + W_{ON} A_{*N} - \frac{1}{c} W_{ON} B_N B_N^T W_{ON} = 0, \tag{28}$$

$$Q_N + A_N^T W_N + W_N A_N - \frac{1}{c} W_N B_N B_N^T W_N = 0. \tag{29}$$

Subtracting (29) from (28), we have

$$2\gamma W_{ON} + \Delta A_N^T W_{ON} + W_{ON} \Delta A_N + A_N^T \Delta W_N + \Delta W_N A_N$$
$$- \frac{1}{c} \Delta W_N B_N B_N^T W_N - \frac{1}{c} W_N B_N B_N^T \Delta W_N - \frac{1}{c} \Delta W_N B_N B_N^T \Delta W_N = 0,$$

or, according to (27),

$$\begin{aligned} &Q_N(W_{ON}) + \left(A_N - \frac{1}{c} B_N B_N^T W_N\right)^T \Delta W_N + \Delta W_N \left(A_N - \frac{1}{c} B_N B_N^T W_N\right) \\ &\quad - \frac{1}{c} \Delta W_N B_N B_N^T \Delta W_N = 0. \end{aligned} \tag{30}$$

The Riccati equation (30) corresponds to an asymptotically stable system with the state and input matrices $A_N - (1/c) B_N B_N^T W_N$ and B_N, respectively. It has a positive definite solution $\Delta W_N > 0$ if $Q_N(W_{ON}) > 0$ [18]. The positive definiteness of $Q_N(W_{ON})$ for all $N \geqq N_0$ guarantees $W_{ON} - W_N > 0$ for all A_i of family (2) in the case $\gamma > 0$, and for all $A_i \neq A_{*i}$ in the case $\gamma = 0$. □

DEFINITION 1. The state matrices A_{*i} $(i = 0, 1, \dots, l)$ of family (2) of system (1) are called the strong dominant matrices if, for the corresponding family of the approximate systems (14) and (2), a number N_0 can be found such that condition (27) is satisfied for all $N \geqq N_0$ and $\gamma = 0$.

DEFINITION 2. The state matrices $A_{*0} + \gamma I$ and $A_{*i}e^{\gamma \tau_i}$ $(i = 1, \dots, l)$ are called the dominant state matrices of family (2) of system (1) if, for the corresponding family of the approximate systems (14) and (2), a number N_0 and $\gamma > 0$ can be found such that condition (27) is satisfied for all $N \geqq N_0$.

According to Theorem 2, the following property of dominant and strong dominant matrices can be established.

For the system (1) ((14)) with strong dominant matrices, the minimal value of the cost functional (3) is more than that for the system with any other matrices of family (2), i.e., $V_0(x(\xi)) - V(x(\xi))$ or $V_{ON} > V_N (W_{ON} > W_N)$ for $N \geqq N_0$.

System (1) with dominant matrices has the minimal value of the cost functional (3), which is more than its minimal value for arbitrary matrices of family (2). Formally, the system with the dominant matrices can be considered as the system with the strong dominant matrices of the extended domain (2), i.e.,

$$\begin{aligned} &A_{1i} \leqq A_i \leqq A_{2i} \quad \text{or} \quad A_i = A_{*i} e^{\gamma \tau_i}, \quad i = 1, \dots, l \\ &A_{10} \leqq A_0 \leqq A_{20} \quad \text{or} \quad A_0 = A_{*0} + \gamma I. \end{aligned} \tag{31}$$

If system (1) with the state matrices has no strong dominant matrices, we can build such matrices by extending domain (2). By choosing certain A_{*i} $(i = 0, 1, \ldots, l)$ of family (2) and by including the matrices $A_{*0} + \gamma I$ and $A_{*i}e^{\gamma\tau_i}$ $(i = 1, \ldots, l)$ in the extended domain (see (31)), we can expect that for some $\gamma > 0$ condition (27) will be satisfied.

Remark. It is more attractive to introduce the dominant and the strong dominant matrices not on the basis of condition (27) but on the basis of their established property. This would widen a class of dominant matrices. Moreover, the existence of the dominant matrices of type $A_{*0} + \gamma I$ and $A_{*i}e^{\gamma\tau_i}$ $(i = 1, \ldots, l)$ would follow from the fact that the minimal value of (13) increases with an increase of γ. However, because the established condition (27) is the sufficient condition of the existence of the strong dominant matrices and the dominant matrices, the definitions of these matrices are linked with this condition. By analyzing (27) we can expect that for large γ the first term of (27) will be dominant so that $Q_N(W_{ON})$ will be positive definite. This would mean that there always exists the possibility to choose the proper γ from the sufficient condition (27). However, this problem needs special consideration.

THEOREM 3. *Suppose system* (1) *with the elements of the state matrices* A_i $(i = 0, 1, \ldots, l)$ *satisfying* (2) *has strong dominant matrices* A_{*i} $(i = 0, 1, \ldots, l)$. *Then, the optimal control law* (4), *which minimizes the cost functional* (3) *subject to system* (1) *with these matrices, makes the system* (1) *and* (4) *asymptotically stable for all matrices* A_i $(i = 0, 1, \ldots, l)$ *satisfying* (2).

Proof. For arbitrary A_i $(i = 0, 1, \ldots, l)$ of family (2) there exists the optimal control (4) and the positive definite solution W, $B(\xi)$, and $P(\xi, \sigma)$ of the Riccati equations (5)–(10) (the quantities corresponding to the strong dominant matrices are denoted by the lower index "0"). Consider the system with arbitrary state matrices A_i and the control (4) determined for the strong dominant matrices

$$\dot{x}(t) = \sum_{i=0}^{l} A_i x(t - \tau_i) - \frac{1}{c} BB^T\left(W_0 x(t) + \int_{-\tau}^{0} B_0(\xi)x(t + \xi)\, d\xi\right). \tag{32}$$

Its finite approximation has the form

$$\dot{x}_N(t) = A_N x_N(t) - \frac{1}{c} B_N B_N^T W_{ON} x_N(t). \tag{33}$$

The derivative of the positive definite form $V_{ON} = 1/2x_N^T(t)W_{ON}x_N(t)$ along (33) is

$$\begin{aligned} 2\frac{dV_{ON}}{dt} &= x_N^T\left(A_N^T W_{ON} + W_{ON}A_N - \frac{2}{c} W_{ON}B_N B_N^T W_{ON}\right)x_N \\ &= x_N^T\left(A_{0N}^T W_{ON} + W_{ON}A_{0N} - \Delta A_N^T W_{ON} - W_{ON}\Delta A_N - \frac{2}{c} W_{ON}B_N B_N^T W_{ON}\right)x_N \\ &= x_N^T\left(-Q_N - \frac{1}{c} W_{ON}B_N B_N^T W_{ON} - (\Delta A_N^T W_{ON} + W_{ON}\Delta A_N)\right)x_N < 0. \end{aligned} \tag{34}$$

Its negative definiteness follows immediately from the nonnegative definiteness of $\Delta A_N^T W_{ON} + W_{ON}\Delta A_N$ (see (27)). Hence, system (33) is asymptotically stable for all state matrices A_i $(i = 0, 1, \ldots, l)$ of family (2). Theorem 1 establishes the proximity of behavior of the approximate and original closed-loop optimal systems (14) and (1). Analogous to the proof of Theorem 1, the proximity of the solutions of (32) and (33) can be established. Therefore, system (32) is asymptotically stable for all matrices A_i ($i = 0, 1, \dots, l$) of (2). □

THEOREM 4. *Suppose system* (1) *with uncertain parameters* (2) *has dominant state matrices* $A_{*0} + \gamma I$ *and* $A_{*i}e^{\gamma\tau_i}$, *where* A_{*i} ($i = 0, 1, \dots, l$) *are fixed matrices of family* (2). *Then, optimal control law* (12) *that minimizes the cost functional* (13) *subject to system* (1) *with the fixed* A_{*i} ($i = 0, 1, \dots, l$) *makes the system* (1) *and* (12) *asymptotically stable for all state matrices* A_i *satisfying* (2).

Proof. The minimization of the functional (13) subject to (1) is equivalent to the minimization of (3) subject to the system with state matrices $A_{*0} + \gamma I$ and $A_{*i}e^{\gamma\tau_i}$ (see Appendix 1). The existence of dominant state matrices is equivalent to the existence of the strong dominant matrices of type $A_{*0} + \gamma I$ and $A_{*i}e^{\gamma\tau_i}$ for system (1) with uncertain parameters belonging to domain (31). The needed result follows immediately from Theorem 3. □

The problem of robust control of system (1) is solved in two steps. First, an estimate of the appropriate γ should be established. Then, the optimal control problem (13) for system (1) with fixed constant parameters is considered. The chosen state dominant matrices should satisfy (27).

As an estimate of γ, the upper bound to the real parts of the eigenvalues of (1) of family (2) can be used, i.e., such a number that for any matrices of A_i ($i = 0, 1, \dots, l$) satisfying (2) the zeros of the characteristic determinant $\det(sI - \sum_{i=0}^{l} A_i e^{-\tau_i s})$ are in the half plane $\{s\colon \operatorname{Re} s \leqq \gamma\}$ (see Appendix 2).

$$\gamma = \max\left[1, (m + \varepsilon) \max_{\substack{r, a_{ikj} \\ \operatorname{Re} s \geqq 0}} |M_r|\right] \tag{35a}$$

$$\gamma = \min\left[\max_{k, a_{ikj}} \sum_{j=1}^{m} \sum_{i=0}^{l} |a_{ikj}|, \max_{j, a_{ikj}} \sum_{j=1}^{m} \sum_{i=0}^{l} |a_{ikj}|\right] \tag{35b}$$

$$\gamma = \frac{1}{2} m \max_{k, j, a_{ikj}} \sum_{i=0}^{l} |a_{ikj} + a_{ijk}|, \tag{35c}$$

where $M_r(s)$ denotes the sum of all the principle minors of order r, $1 \leqq r \leqq m$ of the matrix $\sum_{i=0}^{l} A_i e^{-\tau_i s}$; ε is a small positive number.

Now we assume that the elements of the input matrix B belong to a domain, which is described by

$$B_1 \leqq B \leqq B_2, \tag{36}$$

where, unlike (2), the upper and lower bounds of b_{kj} are assumed to have the same sign.

Such a condition corresponds to the realistic case when an investigator knows qualitatively (but not quantitatively) how the controls influence the state variables.

THEOREM 5. *Let the control* (12) *be determined for the functional* (13), *arbitrary state matrices of the form* (2), *which correspond to the dominant state matrices* $A_{*0} + \gamma I$, $A_{*i}e^{\gamma\tau_i}$ *and the input matrix equal to* B_1. *Then, the closed-loop system* (1) *and* (12) *are asymptotically stable for all the matrices* A_i ($i = 0, 1, \dots, l$) *and* B *of the form* (2) *and* (36), *respectively.*

This theorem can be considered as part of the more general theorem that will be given below.

Let us consider the nonlinear controllable plant of the type

$$\dot{x}(t) = \sum_{i=0}^{l} A_i x(t-\tau_i) + Bg(u(t)), \tag{37}$$

where A_i and B are matrices of the form (2) and (36); $g(u) = (g(u_1), \ldots, g(u_n))$ is a vector function whose elements satisfy the conditions

$$h_{ii}^0 u_i^2 \leqq u_i g_i(u_i), \quad h_{ii}^0 > 0 \quad (i = 1, \ldots, n). \tag{38}$$

(It is assumed that $g(u)$ is smooth enough to guarantee the existence of the solution of (37) for any given initial conditions and controls u.)

THEOREM 6. *Let there exist the optimal solution* (12) *of the problem* (1) *and* (13) *for* A_i *of the form* (2) *that correspond to the dominant state matrices* $A_{*0} + \gamma I$, $A_{*i}e^{\gamma\tau_i}$, *and* $B = B_1 H$, $H^0 - H \geqq 0$, $H = (h_{ii}) > 0$, *and* $H^0 = (h_{ii}^0) > 0$. *Then the nonlinear system* (12) *and* (37) *is absolutely stable for all* A_i *and* B *of the form* (2) *and* (36), *respectively.*

For any fixed A_i and B, the control (4) of system (1) with the input matrix BH, which is optimal with respect to the performance index (3), or the control (12), which is optimal with respect to the performance index (13), makes the nonlinear system (37) asymptotically stable (see the corresponding theorem in [21] and Appendix 3).

In the case of B belonging to the domain (36) we can describe the term $Bg(u)$ by introducing a set of nonlinear functions $\bar{g}(u)$ satisfying (38) and the condition $Bg(u) = B_1\bar{g}(u)$. This is due to the fact that only the lower bound is used in (38).

Hence, instead of system (37) with the family of input matrices B and the given $g(u)$, we can examine the nonlinear system with the constant matrix B_1

$$\dot{x}(t) = \sum_{i=0}^{l} A_i x(t-\tau_i) + B_1\bar{g}(u), \tag{39}$$

where the vector function $\bar{g}(u)$ satisfies condition (38).

According to the theorem mentioned above, the system (37) and (12) under the control determined for $B = B_1 H$ is asymptotically stable.

Finally, according to Theorem 4, the optimal control (12) determined for system (1) with fixed A_i and $B = B_1 H$ makes the system (1) and (12) asymptotically stable for the entire family of A_i and B $(i = 0, 1, \ldots, l)$. □

Example. First we consider the problem of robust control for the system

$$\dot{x}_1 = a_{11}x_1 + u,$$

$$\dot{x}_2 = a_{22}x_2 + u,$$

where

$$-1 \leqq a_{11} \leqq 0, \ a_{22} = 1.$$

According to (35), we can choose $\gamma = 2$.

Let

$$A_{*0} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

To prove that $A_{*0} + 2I$ is the dominant matrix, we will solve the optimal problem with $q_{ii} = 1$, $q_{ij} = 0$, $i \neq j$, and $c = 1000$ (see (13)) and then use condition (27) of

Theorem 2. The solution W_O of the algebraic Riccati equation of type (28) is

$$W_O = \begin{bmatrix} 1 & -1.2 \\ -1.2 & 1.501 \end{bmatrix} \cdot 10^5.$$

Condition (27) has the following form:

$$4 \cdot 10^5 \begin{bmatrix} 1 + 0.5\Delta a_{11} & -1.2(1 + 0.25\Delta a_{11}) \\ -1.2(1 + 0.25\Delta a_{11}) & 1.501 \end{bmatrix} > 0,$$

where $0 \leqq \Delta a_{11} \leqq 1$.

It is easy to check the positive definiteness of this matrix. The robust control law follows from expression (12):

$$u = -20x_1 + 30x_2.$$

Let the first equation of the system examined contain an additional term with delay. We consider the problem of the analytical design of a controller for the system

$$\dot{x}_1 = a_{11}x_1 + 0.05x_1(t - 0.01) + u,$$

$$\dot{x}_2 = x_{22} + u,$$

where $-1 \leqq a_{11} \leqq 0$.

The approximate solutions of optimal control problem (13) for

$$\gamma = 2, \quad A_{*1} = \begin{bmatrix} 0.05 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad A_{*0} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

showed little change between the first- and second-order solutions (i.e., $N = 1$ and $N = 2$), and the computations were terminated at $N = 2$ (this is stipulated by the smallness of the delayed term). Here we give the approximate solution which corresponds to $N = 1$:

$$A_N = \begin{bmatrix} 2 & 0.05e^{0.02} & 0 \\ 100 & -98 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad W_N = \begin{bmatrix} 1.17 & 6 \cdot 10^{-4} & -1.38 \\ 6 \cdot 10^{-4} & 10^{-5} & -7 \cdot 10^{-4} \\ -1.38 & -7 \cdot 10^{-4} & 1.7 \end{bmatrix} \cdot 10^5.$$

As done before, we can see that condition (27) is satisfied and the robust control law is

$$u(t) = -21.8x_1 - 0.01x_1(t - 0.01) + 31.9x_2.$$

Conclusions. The proposed approach of designing a class of robust control systems is based on the consideration of the special optimal control problem for the system with the specified constant parameters. The given procedure allows us to build robust linear systems, as well as a wide class of robust nonlinear systems.

Appendix 1. Let the control law (4) minimize the functional (3) subject to system (1). Consider the problem of minimizing the functional (13) subject to system (1). By introducing

$$x_\gamma(t) = x(t)e^{\gamma t}, \qquad u_\gamma(t) = u(t)e^{\gamma t},$$

the functional (13) becomes of the form (3) with respect to $x_\gamma(t)$ and $u_\gamma(t)$, and system (1) is transformed into

$$\dot{x}_\gamma(t) = (A_0 + \gamma 1)x_\gamma(t) + \sum_{i=1}^{l} A_i e^{\gamma\tau_i} x_\gamma(t - \tau_i) + Bu_\gamma(t).$$

Hence, the problem of minimizing the functional (13) subject to system (1) is equivalent to the usual problem of minimizing the functional of the form (3) subject to the system given above, and the optimal control law $u_\gamma(t)$ can be written as

$$u_\gamma(t) = u(t)e^{\gamma t} = -\frac{1}{c}B^T\left[Wx(t)e^{\gamma t} + \int_{-\tau}^{0} B(\xi)e^{\gamma t}e^{\gamma\xi}x(t+\xi)\,d\xi\right],$$

i.e., $u(t)$ has the form (12).

Appendix 2. The characteristic determinant $f(s)$ of system (1) is

$$f(s) = \det\left(sI - \sum_{i=0}^{l} A_i e^{-\tau_i s}\right) = s^m + P_1(s)s^{m-1} + \cdots + P_m(s),$$

where

$$P_r(s) = (-1)^r M_r(s),$$

and $M_r(s)$ is the sum of all the principal minors of order r of the matrix

$$\sum_{i=0}^{l} A_i e^{-\tau_i s}.$$

The coefficients of $P_r(s)$ belong to some domains, which are defined by the elements of A_i ($i = 0, 1, \ldots, l$) in (1). The location of the eigenvalues of (1) depends on these domains.

We will determine a domain of the right half of the complex plane, Re $s \geqq 0$, which does not contain eigenvalues of the system (1) for all possible a_{Kij} satisfying (2).

From the above expression of $f(s)$, we have

$$|f(s)| \geqq |s|^m\left[1 - \sum_{r=1}^{m} |P_r(s)|\,|s|^{-r}\right]$$

$$\geqq |s|^m\left[1 - \sum_{r=1}^{m} K_{\max}|s|^{-r}\right],$$

where

$$K_{\max} = \max_{\substack{r, a_{iKj} \\ \operatorname{Re} s \geqq 0}} |M_r(s)|.$$

Assuming $|s| \geqq 1$, we have

$$|f(s)| \geqq |s|^m\left[1 - \frac{mK_{\max}}{|s|}\right] = |s|^{m-1}(|s| - mK_{\max}).$$

Hence,

$$|f(s)| > 0 \quad \text{if} \quad |s| \geqq \max\,[1, (m+\varepsilon)K_{\max}],$$

where ε is a small positive number.

The upper bound (35a) of the eigenvalues of (1) follows immediately from this expression.

LEMMA. *If* $\sum_{i=0}^{l} A_i e^{-s\tau_i}$ *is a complex mxm matrix* $A(s) = [a_{kj}(s)]$ *and* $|a_{kk}(s)| > \sum_{j \neq K}^{m} |a_{Kj}(s)|$ *for all* Re $s \geqq 0$, *then* det $|A(s)| \neq 0$ *in* Re $s \geqq 0$. The proof is analogous to the proof of the Levy–Desplanques theorem [13].

We will use this lemma to obtain the estimate (35b). Let s_0 be an eigenvalue of (1) with Re $s_0 \geqq 0$. Then

$$\det |s_0 I - A(s)| = 0$$

and, according to the lemma, at least for one k

$$\left| s_0 - \sum_{i=0}^{l} a_{ikk} e^{-\tau_i s_0} \right| \leqq \sum_{j \neq K}^{l} \left| \sum_{i=0}^{l} a_{ikj} e^{-\tau_i s_0} \right|$$

and

$$|s_0| \leqq \max_{\operatorname{Re} s \geqq 0} \sum_{j=1}^{m} \sum_{i=0}^{l} \left| a_{ikj} e^{-\tau_i s_0} \right| \leqq \max_{\substack{\operatorname{Re} s \geqq 0 \\ k}} \sum_{j=1}^{m} \sum_{i=0}^{l} \left| a_{ikj} e^{-\tau_i s_0} \right|$$

$$\leqq \max_{k} \sum_{j=1}^{m} \sum_{i=0}^{l} |a_{ikj}|.$$

Analogously, considering $A^T(s)$, we obtain

$$|s_0| \leqq \max_{j} \sum_{k=1}^{m} \sum_{i=0}^{l} |a_{ikj}|.$$

Finally, inequality (35c) can be obtained by using the procedure proposed by Hirsch [13]. Let x be a unit eigenvector that corresponds to an eigenvalue s_0, i.e., $A(s_0)x = s_0 x$ and the scalar product $(x, x) = 1$. Then, $(A(s_0)x, x) = s_0$, $(A^*(\bar{s}_0)x, x) = (x, A(s_0)x) = \bar{s}_0$, where the symbol "*" denotes the transpose complex-conjugated matrix and the bar "¯" denotes a complex conjugated value.

Hence,

$$\operatorname{Re} s_0 = \frac{1}{2}\left[(A(s_0)x, x) + (A^*(\bar{s}_0)x, x)\right] = \left(\frac{A(s_0) + A^*(\bar{s}_0)}{2} x, x\right)$$

$$= (B(s_0)x, x)$$

and in the right complex half plane

$$\operatorname{Re} s \leqq \sum_{i,j=1}^{m} |b_{ij}(s_0)|\,|x_i|\,|x_j| \leqq \gamma,$$

where γ is described by (35c).

Appendix 3. Consider a positive definite functional

$$2V(x(t+\xi)) = x^T(t) W x(t) + \int_{-\tau}^{0} \left(x^T(t) B(\xi) x(t+\xi) + x^T(t+\xi) B^T(\xi) x(t)\right) d\xi$$

$$+ \int_{-\tau}^{0} \int_{-\tau}^{0} x^T(t+\xi) P(\xi, \sigma) x(t+\sigma)\, d\sigma\, d\xi,$$

where $W = W^T$, $B(\xi)$, and $P(\xi, \sigma)$ are matrices of appropriate dimensions.

We will use the functional $V(x(t+\xi))$ to examine the stability of the nonlinear control system, which is formed by the plant (37) and the controller (4) determined by solving the auxiliary optimal problems (1) and (3). Its derivative dV_n/dt along the equations of the nonlinear system has the form [20], [21]

$$\frac{dV_n}{dt} = \frac{dV_e}{dt} + cu^T(t)u(t) + \left(x^T(t)W + \int_{-\tau}^{0} x^T(t+\xi)B^T(\xi)\,d\xi\right)Bg(u),$$

where dV_e/dt denotes the derivative of V along the equation of the linear system (1) with the input matrix BH and the corresponding control (4). According to (4) and (38), we have

$$\begin{aligned} &cu^T(t)u(t) + \left(x^T(t)W + \int_{-\tau}^{0} x^T(t+\xi)B^T(\xi)\,d\xi\right)Bg(u) \\ &\quad = cu^T(t)u(t) - cu^T H^{-1} g(u) \\ &\quad \leqq cu^T(t)u(t) - cu^T(t)H^0 H^{-1} u(t). \end{aligned}$$

If $H^0 - H \geqq 0$, $H > 0$, then the right part of the inequality is not positive so that $(dV_n/dt) \leqq (dV_e/dt)$ and, hence, the nonlinear system is absolutely stable.

The case of the control (12), which is determined from the condition of the minimum of (13), has no principal differences from the case considered above.

REFERENCES

[1] A. A. Abdul-Wahab, *Robustness measure bounds for optimal model matching control design*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 1178–1180.

[2] B. D. O. Anderson and E. I. Jury, *On robust Hurwitz polynomials*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 1001–1008.

[3] N. B. Argoun, *Frequency domain conditions for stability of perturbed polynomials*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 913–916.

[4] B. R. Barmish, *Invariance of strict Hurwitz property for polynomial with perturbed coefficients*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 935–936.

[5] ———, *A generalization of Kharitonov's four polynomial concept for robust stability problems with linearly dependent coefficient perturbations*, Proc. 1988 Automat. Control Conf., Atlanta, GA, 3 (1988), pp. 1869–1875.

[6] B. R. Barmish and Z. Shi, *Robust stability of perturbed systems with time delays*, Automatica, 25, (1989), pp. 371–387.

[7] A. C. Bartlett, C. V. Hollot, and L. Huang, *Root locations for an entire polytope of polynomials: It suffices to check the edges*, Math. Control Signals and Systems, 1 (1987), pp. 61–71.

[8] D. S. Bernstein and W. M. Haddad, *Robust stability and performance analysis for state-space systems via quadratic Lyapunov bounds*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 239–271.

[9] R. M. Biernacki, H. Hwang, and H. Bhattacharyya, *Robust stability with structural real parameter perturbation*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 495–506.

[10] J. Cieslki, *On possibilities of the extension of Kharitonov's stability test for interval polynomials to the discrete-time case*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 237–238.

[11] M. Fu, A. W. Olbrot, and M. Polis, *Robust stability for time-delay systems: The edge theorem and grafical tests*, Trans. IEEE Automat. Control, 34 (1989), pp. 813–821.

[12] V. L. Kharitonov, *Asymptotic stability of a family of linear differential equations*, Differentsial'nye Uravneniya, 14 (1978), pp. 2086–2088.

[13] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[14] R. V. Patel, M. Toda, and B. Shridar, *Robustness of linear quadratic state feedback designs in the presence of system uncertainty*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 945–949.

[15] I. R. Peterson and C. V. Hollot, *A Riccati equation approach to the stabilization of uncertain systems*, Automatica, 22 (1986), pp. 397–411.

[16] D. D. Siljak, *Polytopes of nonnegative polynomials*, Proc. 1989 Amer. Control Conf., Pittsburgh, PA, 1 (1989), pp. 193–199.

[17] C. B. Soh, C. S. Berger, and K. P. Dabre, *On the stability properties of polynomials with perturbed coefficients*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1033–1036.

[18] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, Trans. Automat. Control, AC16 (1971), pp. 621–634.

[19] R. T. Yanushevsky, *Optimal control of linear differential-difference systems of neutral type*, Internat. J. Control, 49 (1989), pp. 1835–1850.

[20] ———, *Control of Plants with Time-Lag*, Nauka, Moscow, 1978.

[21] ———, *Synthesis of a certain class of absolutely stable nonlinear control systems*, Eng. Cybernetics, (1979), pp. 128–132.

[22] K. M. Zhou and P. P. Khargonekar, *Stability robustness bounds for linear state space models with structured uncertainty*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 621–623.

[23] ———, *Robust stabilization of linear systems with non-bounded time-varying uncertainty*, Systems & Control Lett. 10 (1988), pp. 17–20.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 991-1015, October 1993

006

CONSTRUCTIVE HEURISTICS AND LOWER BOUNDS FOR GRAPH PARTITIONING BASED ON A PRINCIPAL-COMPONENTS APPROXIMATION*

K. S. ARUN† AND V. B. RAO†

Abstract. This paper addresses the problem of partitioning the vertex set of an edge-weighted undirected graph into two parts of *specified* sizes so as to minimize the sum of the weights on edges joining vertices in different parts. This problem is NP-hard and has several important applications in which the graph size is typically large and the brute-force approach (of listing all feasible partitions and comparing costs) is computationally prohibitive. In this paper a new class of algorithms is developed on the basis of a transformation of the graph problem to a geometric problem of clustering a set of points in Euclidean space. Instead of searching through all feasible partitions that meet the size specifications, it is shown that the search can be confined to a set of $n^{p(p+1)/2}$ partitions, where n is the number of vertices in the graph and p is the rank of the $n \times n$ graph connection matrix. Procedures are developed for constructing all such partitions in $O(n^{p(p+3)/2})$ time. For matrices with large ranks, approximation of the connection matrix by a matrix of low rank by using principal-components analysis is suggested. The algorithms presented in this paper are *constructive* algorithms in that they generate a feasible partition of the graph directly from the connection matrix, as opposed to *iterative* algorithms that start from a feasible partition as an initial guess. The algorithms also bound the difference between the achieved cost and the lowest achievable cost. These lower bounds on the cost of the optimal partition are proved to be superior to the Donath–Hoffman lower bound for two significant special cases wherein either the two part sizes are equal or the graph connection matrix has equal row sums. Simulation results on randomly constructed graphs of different sizes clearly demonstrate the effectiveness of these heuristics in terms of both the cost of the constructed partition and the lower bound provided.

Key words. graph partitioning, geometric clustering, rank reduction, eigendecomposition, principal components, lower bounds, strongly separated partitions

AMS subject classifications. 68R10, 90C27, 62H25, 62H30

1. Introduction. The problem addressed in this paper is the following: Given a weighted undirected graph on n vertices and two positive integers n_1 and n_2 such that $n_1 + n_2 = n$, partition the vertex set into two disjoint subsets of cardinality n_1 and n_2 such that the sum of the weights on edges joining vertices in different subsets is minimum.

It is well known that the graph partitioning (GP) problem is NP-hard, i.e., its decision version is NP-complete [1]. A naive brute-force approach to solving the GP problem is to generate all feasible partitions and to select the one with the lowest cost. It is easily seen that the number of feasible partitions and hence the number of computations required by this brute-force algorithm grows *exponentially* with graph size. For example, the number of bisections (partitions with $n_1 = n_2 = \frac{n}{2}$) in a graph on n vertices is $\frac{1}{2}\binom{n}{n/2}$. The GP problem arises in standard cell placement and floor planning in VLSI circuit design [2], [3], in paging of computer programs, and in computer logic partitioning. In many of these applications the graph size is very large, typically around 1000, and the brute-force algorithm for bisectioning requires over 10^{300} computations.

Let n_1 and n_2 be any two positive integers, and let $n \triangleq n_1 + n_2$. Then a *feasible partition* of the set $\mathbf{V} \triangleq \{1, 2, \ldots, n\}$ is a pair of sets (S_1, S_2) such that

$$S_1 \cup S_2 = \mathbf{V}, \quad S_1 \cap S_2 = \emptyset, \quad |S_1| = n_1, \quad |S_2| = n_2.$$

* Received by the editors February 14, 1991; accepted for publication (in revised form) February 24, 1992. This work was partially supported by Strategic Defense Initiative Office of Innovative Science and Technology under contract DAAL03-91-G-0118 managed by the U.S. Army Research Office.

† Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1101 W. Springfield Avenue, Urbana, Illinois 61801 (arun@uicsl.csl.uiuc.edu).

Every feasible partition (S_1, S_2) can be uniquely represented by an *indicator vector* $\underline{s} \in \mathbb{R}^n$ with ith entry defined as

$$s(i) \triangleq \begin{cases} 1 & \text{if } i \in S_1, \\ 0 & \text{if } i \in S_2. \end{cases}$$

For any subset $\mathbf{V}' \subset \mathbf{V}$ define $\text{IND}(\mathbf{V}') \in \mathbb{R}^n$ as

$$\text{IND}_i(\mathbf{V}') \triangleq \begin{cases} 1 & \text{if } i \in \mathbf{V}', \\ 0 & \text{otherwise.} \end{cases}$$

Then the indicator vector $\underline{s}$ for a partition may be constructed as $\underline{s} = \text{IND}(S_1)$. Conversely, the partition may be reconstructed from its indicator vector $\underline{s}$ as $S_1 = \{i \in \mathbf{V} : s(i) = 1\}$ and $S_2 = \{i \in \mathbf{V} : s(i) = 0\}$. We will use Γ to denote the set of all indicator vectors corresponding to feasible partitions of $\mathbf{V}$.

Let $\mathbf{V}$ be the vertex set of an undirected, edge-weighted, connected graph on n vertices with an $n \times n$ *connection matrix* A whose ijth entry $a_{ij} \geqq 0$ denotes the weight on the edge $\{i, j\}$ in the graph. We assume $a_{ij} = 0$ if there is no edge between vertices i and j in the graph. The connection matrix A serves as a representation of the graph. Clearly, A is a symmetric matrix since the graph is undirected. Define $\underline{1} = (11 \cdots 1)^t$ as the $n \times 1$ vector of ones. The *cost* of a feasible partition $\underline{s} \in \Gamma$ with respect to A is defined to be

$$\text{cost}_A(\underline{s}) \triangleq \underline{s}^t A(\underline{1} - \underline{s}) = \sum_{i \in S_1} \sum_{j \in S_2} a_{ij}. \tag{1.1}$$

The GP problem on A addressed in this paper may be formulated as follows: Given a symmetric $n \times n$ matrix A, find a feasible partition $\underline{s}^* \in \Gamma$ such that $\text{cost}_A(\underline{s}^*) \leqq \text{cost}_A(\underline{s})$ for all $\underline{s} \in \Gamma$. A special case of the GP problem is the *graph bisectioning* (GB) problem, wherein n is even and $n_1 = n_2 = n/2$.

Our approach to solving the GP problem on the connection matrix A of a graph is to approximate A by a simpler matrix of small rank by using principal-components analysis [4], so that GP on the small-rank matrix can be solved exactly in polynomial time. Motivation for such an approximation is based on an observation in [5] that there exists a rank-one, positive semidefinite matrix Q such that a solution to the GP problem on Q is also a solution on A. The principal-components approach is computationally attractive because of the availability of Procedure PART-mD, described in § 4, which allows the computation of an exact solution to the GP problem on a positive semidefinite matrix of rank $= p$ in $O(n^{p(p+3)/2})$ time. Our approach does not exploit any special structure (apart from the connection matrix's possible proximity to low rank) that the graph might possess, such as planarity or near planarity [6].

The principal-components approach was first motivated by the work of Frankle and Karp [7], whose heuristic involved sorting an eigenvector of the graph's connection matrix. In brief, the principal-components approach involves projecting the graph's connection matrix onto the orthogonal complement of Span $(\underline{1})$, computing a set number of principal eigenvalues and eigenvectors of the projected matrix, and solving the GP problem exactly on the low-rank matrix obtained from the principal eigencomponents. Our presentation here includes the derivation of an algorithm for exact solution of the GP problem on low-rank matrices. This approach based on rank-one approximation was first proposed in [5]. An algorithm for exact solution of the GP problem on rank-two matrices was first presented in [8], and the principal-components approach with rank-two approximation was suggested there. The general low-rank case was presented first in

[9], [10]. The principal-components approach also allows for easy computation of lower bounds on the best partition by using the eigencomponents left out of the approximation. The approach provides a degree of freedom in choosing diagonal entries that may be added to the connection matrix before the projection on Span $(\underline{1})$. Experiments with two different choices of diagonal entries were reported in [8].

There are interesting connections between the work here and some work on Euclidean distance matrices. A symmetric matrix with only zeros on the main diagonal is said to be a Euclidean distance matrix embedded in q dimensions if there exist n points in a q-dimensional Euclidean space such that the (i, j)th entry of the matrix is the square of the distance between the ith and jth points. It can be shown that a graph connection matrix is a Euclidean distance matrix embedded in q dimensions if and only if the matrix has rank q after projection onto the orthogonal complement of Span $(\underline{1})$ [11], [12].

An algorithm for graph partitioning based on eigendecomposition was recently proposed by Pothen, Simon, and Liou [13]. It turns out that their algorithm is identical to one of the principal-components heuristics studied here, specifically, the principal-components algorithm based on rank-one approximation and equal-row-sum choice of diagonal entries. The approach expounded here is more general in that it allows for arbitrary choice of the diagonal entries and higher-rank approximations. It also leads to tighter lower bounds.

The rest of this paper is organized as follows. Section 2 presents a transformation of the GP problem to a geometric problem of clustering a set of points in Euclidean space. In §§ 3 and 4 algorithms are developed for solving the GP problem on rank-one and low-rank matrices exactly in polynomial time. Section 5 provides a means for constructing low-rank positive semidefinite approximations to the connection matrix via principal-components analysis. Heuristics for solving the GP problem on connection matrices of large rank are presented in § 6. Methods for improving the approximations of § 5 by using clever choices of diagonal entries of the connection matrix are studied in § 7. New lower bounds on the cost of any feasible partition are presented in § 8. The paper concludes with a presentation of simulation results on graphs of different sizes in § 9.

2. A problem transformation. We begin by noting that the diagonal entries of A do not affect $\mathrm{cost}_A(\underline{s})$ defined in (1.1). Therefore, for *any* diagonal matrix D

$$\mathrm{cost}_A(\underline{s}) = \mathrm{cost}_{A+D}(\underline{s}) \quad \text{for all} \quad \underline{s} \in \Gamma.$$

This is a useful observation that leads to a transformation of the GP problem into a geometric clustering (GC) problem in Euclidean space, where lower-dimensional approximations can be made. It also provides a degree of freedom in selecting the matrix to be approximated by its principal components. On the basis of the above observation one can construct a positive semidefinite matrix from A by changing only its diagonal entries, without affecting the solution to the GP problem. One approach is to subtract the smallest eigenvalue of A from its diagonal elements. Therefore, without loss of generality we will henceforth assume that A is positive semidefinite. Being positive semidefinite, the A matrix admits the following square-root factorization: $A = B^t B$, where B is a $k \times n$ matrix and $k = \mathrm{rank}\,(A)$. Therefore,

$$\begin{aligned} \mathrm{cost}_A(\underline{s}) &= [B\underline{s}]^t[B(\underline{1} - \underline{s})] = \tfrac{1}{2}(\|B\underline{1}\|^2 - \|B\underline{s}\|^2 - \|B(\underline{1} - \underline{s})\|^2) \\ &= \tfrac{1}{4}(\|B\underline{1}\|^2 - \|B\underline{s} - B(\underline{1} - \underline{s})\|^2). \end{aligned} \tag{2.1}$$

Hence solving the GP problem on A is equivalent to solving the following (GC) problem on B: Given a $k \times n$ matrix B, find $\underline{s} \in \Gamma$ that *maximizes* $\|B\underline{s}\|^2 + \|B(\underline{1} - \underline{s})\|^2$.

The solution to the GC problem on B also maximizes $\|B\underline{\mathbf{s}} - B(\underline{\mathbf{1}} - \underline{\mathbf{s}})\|^2$. If we use $\underline{\mathbf{b}}_j \in \mathbb{R}^k$ to denote the jth column of B, then the GC problem on B reduces to partitioning a set $\{\underline{\mathbf{b}}_1, \underline{\mathbf{b}}_2, \ldots, \underline{\mathbf{b}}_n\}$ of n vectors in $\mathbb{R}^k$ into two sets $\{\underline{\mathbf{b}}_i : i \in S_1\}$ and $\{\underline{\mathbf{b}}_j : j \in S_2\}$ with $|S_1| = n_1$ and $|S_2| = n_2$ such that $\|\sum_{i\in S_1} \underline{\mathbf{b}}_i\|^2 + \|\sum_{j\in S_2} \underline{\mathbf{b}}_j\|^2$ and $\|\sum_{i\in S_1} \underline{\mathbf{b}}_i - \sum_{j\in S_2} \underline{\mathbf{b}}_j\|^2$ are maximized.

The above transformation of the GP problem on A to the GC problem on B was first used by Frankle and Karp [7] for the bisectioning case, and their heuristic for solving the GB problem is briefly discussed in § 5. One of the principal-components-based heuristics of this paper was first developed for the GB problem in [5] as an alternative to the Frankle–Karp approach. Extensions for handling unequal part sizes appear in [8].

It must be mentioned that the GC problem is also NP-hard, primarily because the dimension k of the space in which the problem is formulated grows, in general, with the size n of the problem. The principal-components approach to solving the GC problem is first to approximate the given points in $\mathbb{R}^k$ by their projections on a low-dimensional linear variety and then to solve the simpler GC problem there.

3. An algorithm for the 1D problem. Throughout this paper 1D will be used to denote the term "one-dimensional." Likewise, pD will be used to denote "p-dimensional" for any integer $p \geqq 2$. Also, all matrices will be denoted by italic upper-case letters, and vectors will be denoted by underlined bold lower-case letters. The jth column of a matrix A is $\underline{\mathbf{a}}_j$, and its ijth entry is either a_{ij} or $a_j(i)$. Finally, the ith entry of a vector $\underline{\mathbf{b}}$ is denoted by either b_i or $b(i)$. We begin by demonstrating that solving the GC problem on a $k \times n$ matrix B is computationally simple if all the columns of B lie on a 1D linear variety.

Consider the following algorithm for generating a pair of feasible partitions based on sorting the entries of a vector $\underline{\mathbf{e}} \in \mathbb{R}^n$. In the following procedure $\underline{\mathbf{s}} = \text{SELECT}(\underline{\mathbf{e}}, k)$ is a binary vector of the same size as $\underline{\mathbf{e}}$ and with exactly k entries set to unity and with the other $n - k$ entries set to zero. The indices of the entries set to unity correspond to the k largest entries in $\underline{\mathbf{e}}$. The function SELECT can be implemented in $O(n)$ time [14]. When the entries of $\underline{\mathbf{e}}$ are not all distinct, different implementations of the function SELECT may generate different outputs, depending on the strategy used to break ties. The pair of feasible partitions generated by the following procedure will therefore depend on the implementation of the function SELECT when there is a tie for either the n_1th-largest or the n_2th-largest entry in $\underline{\mathbf{e}}$. Procedure PART-1D can be used to generate an optimal solution to the clustering problem on any set of points that are on a 1D subspace or a 1D linear variety.

```
Procedure PART-1D ($\underline{\mathbf{e}}$, $n_1$, $n_2$, $\underline{\mathbf{s}}_a$, $\underline{\mathbf{s}}_b$)
    Input: A vector $\underline{\mathbf{e}} \in \mathbb{R}^n$ and two positive integers $n_1$ and $n_2$ such that $n_1 + n_2 = n$.
    Output: Two indicator vectors $\underline{\mathbf{s}}_a$ and $\underline{\mathbf{s}}_b$ representing two feasible partitions.
    begin
        $\underline{\mathbf{s}}_a$ = SELECT($\underline{\mathbf{e}}$, $n_1$);
        $\underline{\mathbf{s}}_b$ = NOT(SELECT($\underline{\mathbf{e}}$, $n_2$));
    end
```

PROPOSITION 1. *For any given* $\underline{\mathbf{e}} \in \mathbb{R}^n$, *one of the feasible partitions*, $\underline{\mathbf{s}}_a$ *or* $\underline{\mathbf{s}}_b$, *output by Procedure* PART-1D($\underline{\mathbf{e}}$, n_1, n_2, $\underline{\mathbf{s}}_a$, $\underline{\mathbf{s}}_b$) *is an optimal feasible partition for* GP *on the matrix* $\underline{\mathbf{e}}\underline{\mathbf{e}}^t$.

Proof. From the definition of function SELECT we have

$$\underline{\mathbf{s}}_a^t\underline{\mathbf{e}} \geqq \underline{\mathbf{s}}^t\underline{\mathbf{e}} \geqq \underline{\mathbf{s}}_b^t\underline{\mathbf{e}} \quad \text{for all} \quad \underline{\mathbf{s}} \in \Gamma. \tag{3.1}$$

Define $\alpha \triangleq \underline{\mathbf{s}}_a^t\underline{\mathbf{e}}$, $\beta \triangleq \underline{\mathbf{s}}_b^t\underline{\mathbf{e}}$, $\gamma \triangleq \underline{\mathbf{1}}^t\underline{\mathbf{e}}$, and a function $f(x) \triangleq \gamma x - x^2$ of a real variable x. Since f is strictly concave, its minimum over the interval $[\beta, \alpha]$ is achieved at one of the end points [15]. Hence $f(x) \geqq \min\{f(\alpha), f(\beta)\}$ for all $x \in [\beta, \alpha]$. Replacing x by $\underline{\mathbf{s}}^t\underline{\mathbf{e}}$ and noting that $f(\underline{\mathbf{s}}^t\underline{\mathbf{e}}) = \mathrm{cost}_{\underline{\mathbf{ee}}^t}(\underline{\mathbf{s}})$, we get $\mathrm{cost}_{\underline{\mathbf{ee}}^t}(\underline{\mathbf{s}}) \geqq \min\{\mathrm{cost}_{\underline{\mathbf{ee}}^t}(\underline{\mathbf{s}}_a), \mathrm{cost}_{\underline{\mathbf{ee}}^t}(\underline{\mathbf{s}}_b)\}$ for all $\underline{\mathbf{s}} \in \Gamma$, which proves this proposition. □

COROLLARY. *Given fixed vectors* $\underline{\mathbf{v}}$, $\underline{\mathbf{z}} \in \mathbb{R}^k$ *and* $\underline{\mathbf{e}} \in \mathbb{R}^n$, *where* $k \leqq n$, *define* $\underline{\mathbf{b}}_i = \underline{\mathbf{v}} + e(i)\underline{\mathbf{z}}$ *and a* $k \times n$ *matrix* $B = [\underline{\mathbf{b}}_1, \underline{\mathbf{b}}_2, \dots, \underline{\mathbf{b}}_n]$. *Then one of the feasible partitions* $\underline{\mathbf{s}}_a$ *or* $\underline{\mathbf{s}}_b$ *output by Procedure* PART-1D ($\underline{\mathbf{e}}$, n_1, n_2, $\underline{\mathbf{s}}_a$, $\underline{\mathbf{s}}_b$) *is an optimal feasible partition for* GC *on* B *and, consequently, an optimal partition for* GP *on* $A \triangleq B^tB$.

Proof. First note that the columns of B defined by this corollary lie on a 1D linear variety in $\mathbb{R}^k$ given by $\{\underline{\mathbf{v}} + \alpha\underline{\mathbf{z}} : \alpha \in \mathbb{R}\}$. Let $\underline{\mathbf{v}}_0 = \underline{\mathbf{v}} + \alpha_0\underline{\mathbf{z}}$ be the minimum-norm element [16] of this variety, which clearly satisfies $\underline{\mathbf{z}}^t\underline{\mathbf{v}}_0 = 0$. Since $\underline{\mathbf{b}}_i = \underline{\mathbf{v}}_0 + (e(i) - \alpha_0)\underline{\mathbf{z}}$, we get for each $\underline{\mathbf{s}} \in \Gamma$

$$\mathrm{cost}_A(\underline{\mathbf{s}}) = \left(\sum_{i \in S_1} \underline{\mathbf{b}}_i\right)^t\left(\sum_{j \in S_2} \underline{\mathbf{b}}_j\right) = n_1n_2\|\underline{\mathbf{v}}_0\|^2 + \underline{\mathbf{s}}^T\underline{\mathbf{e}}'\underline{\mathbf{e}}'^t(\underline{\mathbf{1}} - \underline{\mathbf{s}})\|\underline{\mathbf{z}}\|^2, \tag{3.2}$$

where $\underline{\mathbf{e}}' \in \mathbb{R}^n$ is defined by $e'(i) = e(i) - \alpha_0$. From Proposition 1 we see that one of the feasible partitions, $\underline{\mathbf{s}}_a$ or $\underline{\mathbf{s}}_b$, outputted by Procedure PART-1D ($\underline{\mathbf{e}}'$, n_1, n_2, $\underline{\mathbf{s}}_a$, $\underline{\mathbf{s}}_b$) is an optimal feasible partition for GP on $\underline{\mathbf{e}}'\underline{\mathbf{e}}'^t$, which is also optimal for GP on A by (3.2). But the output of function SELECT on the vector $\underline{\mathbf{e}}$ is the same as the output on the vector $\underline{\mathbf{e}}' = \underline{\mathbf{e}} - \alpha_0\underline{\mathbf{1}}$. This proves the corollary. □

It is important to note that when the solution to the GP problem on $\underline{\mathbf{ee}}^t$ is not unique, Procedure PART-1D does not generate *all* the optimal partitions. However, nonuniqueness of the optimal partition is possible only when the points are not all distinct from one another. Since procedure SELECT on $\underline{\mathbf{e}} \in \mathbb{R}^n$ takes $O(n)$ time, the above results show that solving the GC problem for vectors on a 1D linear variety takes $O(n)$ time. It can be shown [5] that there exists a 1D subspace such that the application of PART-1D to the projections of the $\underline{\mathbf{b}}_i$'s on this subspace results in an optimal partition for GP on B^tB. However, the problem of finding such a subspace is itself computationally intractable; hence we resort to heuristics.

One of the heuristics proposed in this paper uses in place of the optimal 1D subspace a 1D linear variety that is closest to all the points $\{\underline{\mathbf{b}}_i : i \in \underline{\mathbf{V}}\}$ in the least-squares sense. Then a simpler clustering problem is solved with respect to the projections $\tilde{\underline{\mathbf{b}}}_i$ of the original points on the chosen 1D variety. Replacing the points by their projections is equivalent to *approximating* the original connection matrix B^tB by a new matrix $\tilde{B}^t\tilde{B}$ of rank no greater than two. However, an optimal partition for GP on the approximated matrix may not be an optimal partition to the original problem. The quality of the approximation can be improved by using higher-dimensional varieties for the approximation. To this end, in the following section we will develop an algorithm for solving the clustering problem in $\mathbb{R}^p$, where p is a small positive integer.

4. An algorithm for clustering in higher dimensions. The corollary to Proposition 1 establishes that the clustering problem for points on a 1D linear variety in $\mathbb{R}^k$ is as simple as the clustering problem for points on a 1D subspace of $\mathbb{R}^k$. More generally, it can be shown that the clustering problem for points on any pD linear variety in $\mathbb{R}^k$, $p \leqq k$, can be solved by clustering an appropriately translated set of points on a pD subspace of $\mathbb{R}^k$. The translation is achieved by subtracting the minimum-norm element of the variety from every point.

We have seen that the clustering problem in $\mathbb{R}$ is simple and that its solution can be obtained by using PART-1D. We will now show that the clustering problem in $\mathbb{R}^p$ is also relatively simple for small values of p. To this end, we introduce the notion of strongly separated partitions.

DEFINITION. A feasible partition (S_1, S_2) (and its corresponding indicator vector $\underline{\mathbf{s}}$) is said to be *strongly separated* with respect to a $k \times n$ matrix $B = [\underline{\mathbf{b}}_1, \underline{\mathbf{b}}_2, \ldots, \underline{\mathbf{b}}_n]$ if there exists a hyperplane in $\mathbb{R}^k$ that separates the points $\{\underline{\mathbf{b}}_i : i \in S_1\}$ from the points $\{\underline{\mathbf{b}}_j : j \in S_2\}$ and if this hyperplane touches the two sets at most at one point common to both sets. More precisely, (S_1, S_2) is a strongly separated partition with respect to B if there exists a *nonzero* vector $\underline{\mathbf{x}} \in \mathbb{R}^k$ and a scalar $\mu \in \mathbb{R}$ such that

$$\text{(4.1a)} \qquad \underline{\mathbf{x}}^t\underline{\mathbf{b}}_i \geqq \mu \quad \text{for all} \quad i \in S_1, \qquad \underline{\mathbf{x}}^t\underline{\mathbf{b}}_j \leqq \mu \quad \text{for all} \quad j \in S_2$$

and

$$\text{(4.1b)} \qquad \underline{\mathbf{x}}^t\underline{\mathbf{b}}_i = \mu = \underline{\mathbf{x}}^t\underline{\mathbf{b}}_j \quad \text{if and only if} \quad \underline{\mathbf{b}}_i = \underline{\mathbf{b}}_j.$$

Observe that when all columns of B are distinct and (S_1, S_2) is a strongly separated partition with respect to B, then the inequalities in (4.1a) become strict, which in turn implies that the convex hulls of $\{\underline{\mathbf{b}}_i : i \in S_1\}$ and of $\{\underline{\mathbf{b}}_i : i \in S_2\}$ do not intersect. A partition satisfying this nonintersecting convex hull property is called an NICH partition in [8]. Note that every NICH partition is a strongly separated partition, but not vice versa.

PROPOSITION 2. *For a given $k \times n$ matrix B, an optimal partition for* GP *on $A = B^tB$ is necessarily a strongly separated partition with respect to B.*

Proof. Let $\underline{\mathbf{s}}^*$ be the indicator vector corresponding to an optimal feasible partition (S_1^*, S_2^*) that solves the GP problem on A. Define

$$\underline{\mathbf{x}} \triangleq B\underline{\mathbf{s}}^* - B(\underline{\mathbf{1}} - \underline{\mathbf{s}}^*) = 2B\underline{\mathbf{s}}^* - B\underline{\mathbf{1}}.$$

The expression for $\text{cost}_{B^tB}(\underline{\mathbf{s}})$ in (2.1) can be manipulated to give

$$\text{(4.2)} \qquad \text{cost}_A(\underline{\mathbf{s}}) - \text{cost}_A(\underline{\mathbf{s}}^*) = \underline{\mathbf{h}}^t\underline{\mathbf{x}} - \|\underline{\mathbf{h}}\|^2,$$

where $\underline{\mathbf{h}} \triangleq B(\underline{\mathbf{s}}^* - \underline{\mathbf{s}})$. Since $\underline{\mathbf{s}}^*$ is optimal, we must have $\underline{\mathbf{h}}^t\underline{\mathbf{x}} \geqq \|\underline{\mathbf{h}}\|^2$. Thus $\underline{\mathbf{x}}$ cannot be $\underline{\mathbf{0}}$ unless $\underline{\mathbf{h}} = \underline{\mathbf{0}}$ for all $\underline{\mathbf{s}} \in \Gamma$, which would mean $B\underline{\mathbf{s}} = B\underline{\mathbf{s}}^*$ for all $\underline{\mathbf{s}} \in \Gamma$. It can be seen that $B\underline{\mathbf{s}}$ can be constant for all $\underline{\mathbf{s}} \in \Gamma$ only if the $\underline{\mathbf{b}}_i$ are all identical. Thus when $\underline{\mathbf{x}} = \underline{\mathbf{0}}$, every partition is optimal, every partition is strongly separated, and the proposition is trivially true.

When $\underline{\mathbf{x}} \neq \underline{\mathbf{0}}$, consider a fixed but arbitrary $i \in S_1^*$ and $j \in S_2^*$. Let $\underline{\mathbf{s}}$ be the indicator vector of the partition (S_1, S_2) obtained by interchanging i and j from the optimal partition. More precisely, $S_1 = S_1^* \cup \{i\} - \{j\}$ and $S_2 = S_2^* \cup \{j\} - \{i\}$. Note that $\underline{\mathbf{h}} = \underline{\mathbf{b}}_i - \underline{\mathbf{b}}_j$ in this case. Since $\underline{\mathbf{s}}^*$ is optimal, (4.2) yields $\underline{\mathbf{h}}^t\underline{\mathbf{x}} \geqq 0$ or $\underline{\mathbf{b}}_i^t\underline{\mathbf{x}} \geqq \underline{\mathbf{b}}_j^t\underline{\mathbf{x}}$ with equality if and only if $\underline{\mathbf{h}} = \underline{\mathbf{0}}$ or $\underline{\mathbf{b}}_i = \underline{\mathbf{b}}_j$. The above is true for every pair (i, j) such that $i \in S_1^*$ and $j \in S_2^*$. Pick a scalar μ such that

$$\min_{i \in S_1^*} \underline{\mathbf{x}}^t\underline{\mathbf{b}}_i \geqq \mu \geqq \max_{j \in S_2^*} \underline{\mathbf{x}}^t\underline{\mathbf{b}}_j,$$

and observe that the chosen $\underline{\mathbf{x}}$ and μ satisfy the three conditions of (4.1). Hence the proposition. □

Example 1. Consider the two partitions depicted in Fig. 1. In each case, there exists a hyperplane that separates the two sets $\{\underline{\mathbf{b}}_i : i \in S_1\}$ and $\{\underline{\mathbf{b}}_i : i \in S_2\}$. But the partition in Fig. 1(a) is not strongly separated because there is no separating hyperplane that passes *only* through a point common to both parts. Also, the partition in Fig. 1(b) is not strongly separated because the two convex hulls intersect even though the points are all distinct in this case. Therefore, both partitions are not optimal since neither is strongly separated.

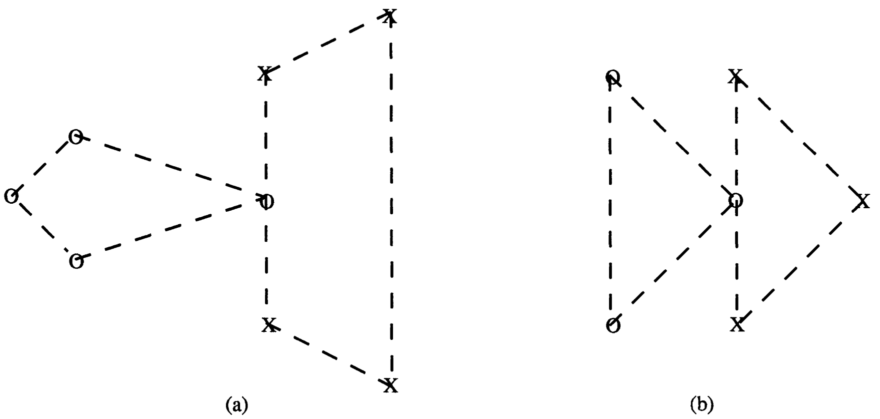


FIG. 1. *Two partitions that are not strongly separated. Points in one set are denoted by* "O," *and those in the other set are denoted by* "X."

Proposition 2 establishes that the GP problem on low-rank, positive semidefinite matrices can be solved relatively easily. The approach it suggests is to work with the points $\underline{b}_i$ in $\mathbb{R}^p$, where p is the rank of the connection matrix, and to construct all feasible, strongly separated partitions. It will be shown below that when the points $\underline{b}_i$ are all distinct, the number of strongly separated feasible partitions is relatively small and manageable. We next describe a polynomial-time algorithm PART-mD that generates a set Φ of strongly separated feasible partitions with respect to a $p \times n$ matrix B that contains an optimal solution to the GP problem on B^tB. The size of this set Φ is at most $n^{p(p+1)/2}$, and the search for an optimal partition can be confined to Φ. When the columns of B are distinct (they may be linearly dependent), the set Φ generated by PART-mD consists of *all* strongly separated feasible partitions with respect to B. When the columns are not distinct, not every strongly separated feasible partition is generated by the procedure, but for every such partition not in Φ there is guaranteed to be one in Φ that has the same cost. The procedure relies on the following proposition, which establishes that strong separation is preserved even when attention is restricted to a subset of points that lie on a lower-dimensional linear variety in $\mathbb{R}^p$.

PROPOSITION 3. *Let* (S_1, S_2) *be a strongly separated partition of* $\mathbf{V} = \{1, 2, \dots, n\}$ *with respect to a* $p \times n$ *matrix* $B = [\underline{b}_1, \underline{b}_2, \dots, \underline{b}_n]$. *Consider a fixed* $\underline{w} \in \mathbb{R}^p$, *an integer* $s < p$, *and a* $p \times s$ *matrix* Z *with orthonormal columns. Let* $W \triangleq \{\underline{w} + Z\underline{e} : \underline{e} \in \mathbb{R}^s\}$ *be an sD linear variety in* $\mathbb{R}^p$, *and define* $\mathbf{J} \triangleq \{i \in \mathbf{V} : \underline{b}_i \in W\}$. *For each* $i \in \mathbf{J}$ *define* $\underline{e}_i \triangleq Z^t(\underline{b}_i - \underline{w})$ *as the representation of* $\underline{b}_i$ *with respect to* Z. *If* $S_1 \cap \mathbf{J}$ *and* $S_2 \cap \mathbf{J}$ *are each nonempty, then* $(S_1 \cap \mathbf{J}, S_2 \cap \mathbf{J})$ *is a strongly separated partition of* $\mathbf{J}$ *with respect to the matrix* $[\underline{e}_i : i \in \mathbf{J}]$.

Proof. First note that for any subset $\mathbf{J}$ of $\mathbf{V}$, $(S_1 \cap \mathbf{J}, S_2 \cap \mathbf{J})$ is a strongly separated partition of $\mathbf{J}$ with respect to B as long as $S_1 \cap \mathbf{J}$ and $S_2 \cap \mathbf{J}$ are both nonempty. Hence, by definition, there exists $\underline{x} \in \mathbb{R}^r$ and a scalar μ such that

$$\underline{x}^t\underline{b}_i \geqq \mu \quad \text{for all} \quad i \in S_1 \cap \mathbf{J}, \qquad \underline{x}^t\underline{b}_j \leqq \mu \quad \text{for all} \quad j \in S_2 \cap \mathbf{J},$$

and

$$\underline{x}^t\underline{b}_i = \mu = \underline{x}^t\underline{b}_j \quad \text{if and only if} \quad \underline{b}_i = \underline{b}_j.$$

But $\underline{x}^t\underline{b}_i = \underline{x}^t\underline{w} + \underline{h}^t\underline{e}_i$ for all $i \in \mathbf{J}$, where $\underline{h} \triangleq Z^t\underline{x}$. Since $\underline{x}^t\underline{w}$ is constant for all $i \in \mathbf{J}$, the vector $\underline{\mathbf{h}}$ and scalar $\mu' = \mu - \underline{\mathbf{x}}^t\underline{\mathbf{w}}$ satisfy the definition of strong separation of the partition $(S_1 \cap \mathbf{J}, S_2 \cap \mathbf{J})$ of $\mathbf{J}$ with respect to $\{\underline{\mathbf{e}}_i : i \in \mathbf{J}\}$. □

This property of strongly separated partitions permits the recursive generation of all such partitions by the Procedure PART-mD. For any hyperplane H in $\mathbb{R}^p$, H^+ and H^- will denote the two strict *half-planes* of $\mathbb{R}^p$ generated by H in this procedure. Points that happen to fall on H are partitioned by a call to Procedure PART-H, which in turn calls PART-mD for a subspace that is one dimension smaller. The procedures are shown below.

```
Procedure PART-mD ($B$, $n_1$, $n_2$, $\boldsymbol{\Phi}$)
    Input: Positive integers $n_1$, $n_2$, and full-rank matrix $B$ with $(n_1 + n_2)$ columns and
    two or more rows.
    Output: A set $\boldsymbol{\Phi}$ of indicator vectors representing strongly separated partitions with
    respect to $B$.
    begin
        $\boldsymbol{\Phi} \leftarrow \emptyset$;
        $p \leftarrow$ Number of rows in $B$;
        FOR each set $\{i_1, i_2, \ldots, i_p\}$ of $p$ distinct positive integers $\leqq n_1 + n_2$, DO
            IF $L \triangleq \{\underline{\mathbf{b}}_{i_2} - \underline{\mathbf{b}}_{i_1}, \underline{\mathbf{b}}_{i_3} - \underline{\mathbf{b}}_{i_1}, \ldots, \underline{\mathbf{b}}_{i_p} - \underline{\mathbf{b}}_{i_1}\}$ is linearly independent THEN
                Generate hyperplane $H$ containing $\{\underline{\mathbf{b}}_{i_1}, \underline{\mathbf{b}}_{i_2}, \ldots, \underline{\mathbf{b}}_{i_p}\}$;
                $\underline{\mathbf{s}}^+ = \mathrm{IND}(\{i \in \mathbf{V} : \underline{\mathbf{b}}_i \in H^+\})$; $n^+ = \underline{\mathbf{1}}^t\underline{\mathbf{s}}^+$;
                $\underline{\mathbf{s}}^- = \mathrm{IND}(\{i \in \mathbf{V} : \underline{\mathbf{b}}_i \in H^-\})$; $n^- = \underline{\mathbf{1}}^t\underline{\mathbf{s}}^-$;
                $\underline{\mathbf{s}}^0 = \mathrm{IND}(\{i \in \mathbf{V} : \underline{\mathbf{b}}_i \in H\})$;
                IF $(n^+ \leqq n_1$ AND $n^- \leqq n_2)$ THEN
                    IF $(n^+ = n_1)$        $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} \cup \{\underline{\mathbf{s}}^+\}$;
                    IF $(n^- = n_2)$        $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} \cup \{\underline{\mathbf{s}}^0 + \underline{\mathbf{s}}^+\}$;
                    IF $(n^+ < n_1$ AND $n^- < n_2)$ THEN
                        PART-H ($B$, $\underline{\mathbf{s}}^0$, $L$, $n_1 - n^+$, $n_2 - n^-$, $\boldsymbol{\Phi}'$);
                        FOR each $\underline{\mathbf{s}}' \in \boldsymbol{\Phi}'$ DO
                            $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} \cup \{\underline{\mathbf{s}}^+ + \underline{\mathbf{s}}'\}$;
                        ENDFOR
                    ENDIF
                ENDIF
                IF $(n^- \leqq n_1$ AND $n^+ \leqq n_2)$ THEN
                    IF $(n^- = n_1)$        $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} \cup \{\underline{\mathbf{s}}^-\}$;
                    IF $(n^+ = n_2)$        $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} \cup \{\underline{\mathbf{s}}^0 + \underline{\mathbf{s}}^-\}$;
                    IF $(n^- < n_1$ AND $n^+ < n_2)$ THEN
                        PART-H ($B$, $\underline{\mathbf{s}}^0$, $L$, $n_1 - n^-$, $n_2 - n^+$, $\boldsymbol{\Phi}'$);
                        FOR each $\underline{\mathbf{s}}' \in \boldsymbol{\Phi}'$ DO
                            $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} \cup \{\underline{\mathbf{s}}^- + \underline{\mathbf{s}}'\}$;
                        ENDFOR
                    ENDIF
                ENDIF
            ENDIF
        ENDFOR
    end

Procedure PART-H ($B$, $\underline{\mathbf{s}}^0$, $L$, $n_a$, $n_b$, $\boldsymbol{\Phi}$)
    Input: Positive integers $n_a$, $n_b$, full-rank matrix $B$ with $(n_a + n_b)$ or more columns
    and two or more rows, binary vector $\underline{\mathbf{s}}^0$ indicating the $n_a + n_b$ columns of $B$ that
    lie on a translation of Span $(L)$.
```

```
Output: A set of $\Phi$ of indicator vectors representing strongly separated partitions
on the hyperplane.
begin
    $\mathbf{V}^0 \leftarrow \{i : s^0(i) = 1\}$;
    Let $f: \{1, 2, \ldots, n_a + n_b\} \rightarrow \mathbf{V}^0$ be any ordering of $\mathbf{V}^0$;
    Obtain an orthonormal basis $Z$ for span $(L)$;
    Let $\underline{\mathbf{e}}'_i = Z^t(\underline{\mathbf{b}}_i - \underline{\mathbf{b}}_{i_1})$ represent $\underline{\mathbf{b}}_i$ with respect to $Z$ for each $i \in \mathbf{V}^0$;
    $B' \leftarrow [\underline{\mathbf{e}}_{f(1)} \ \underline{\mathbf{e}}_{f(2)} \ \cdots \ \underline{\mathbf{e}}_{f(n_0)}]$;
    IF ($B'$ has two or more rows) THEN
        PART-mD $(B', n_a, n_b, \Phi')$;
    ELSE
        PART-1D $(B', n_a, n_b, \underline{\mathbf{s}}_q, \underline{\mathbf{s}}_r)$;
        $\Phi' = \{\underline{\mathbf{s}}_q, \underline{\mathbf{s}}_r\}$;
    ENDIF
    $\Phi \leftarrow \varnothing$;
    FOR each $\underline{\mathbf{s}}' = \Phi'$ DO
        $\mathbf{V}' = \{f(i) : s'(i) = 1\}$;
        $\Phi \leftarrow \Phi \cup \mathrm{IND}(\mathbf{V}')$;
    ENDFOR
end
```

The following proposition establishes that Procedure PART-mD generates all relevant strongly separated partitions with respect to B.

PROPOSITION 4. *If* (S_1, S_2) *is a strongly separated partition with respect to a* $p \times n$ *matrix* $B = [\underline{\mathbf{b}}_1, \underline{\mathbf{b}}_2, \ldots, \underline{\mathbf{b}}_n]$ *of rank* p, *then there exists a hyperplane* H *in* $\mathbb{R}^p$ *that passes through a subset* $\{\underline{\mathbf{b}}_{i_1}, \underline{\mathbf{b}}_{i_2}, \ldots, \underline{\mathbf{b}}_{i_p}\}$ *of* p *columns of* B *such that* $\{\underline{\mathbf{b}}_{i_2} - \underline{\mathbf{b}}_{i_1}, \underline{\mathbf{b}}_{i_3} - \underline{\mathbf{b}}_{i_1}, \ldots, \underline{\mathbf{b}}_{i_p} - \underline{\mathbf{b}}_{i_1}\}$ *is linearly independent and the two sets* $\{\underline{\mathbf{b}}_i, i \in S_1\}$ *and* $\{\underline{\mathbf{b}}_i, i \in S_2\}$ *are on opposite half-planes of* H. *More precisely, there exists a nonzero vector* $\underline{\mathbf{x}} \in \mathbb{R}^p$ *and a scalar* μ *such that* $\underline{\mathbf{b}}_i^t\underline{\mathbf{x}} \geqq \mu$ *for all* $i \in S_1$, $\underline{\mathbf{b}}_j^t\underline{\mathbf{x}} \leqq \mu$ *for all* $j \in S_2$, *and* $\underline{\mathbf{b}}_i^t\underline{\mathbf{x}} = \mu$ *for at least* p *indices* $\{i_1, i_2, \ldots, i_p\}$ *for which the set of vectors* $\{\underline{\mathbf{b}}_{i_2} - \underline{\mathbf{b}}_{i_1}, \underline{\mathbf{b}}_{i_3} - \underline{\mathbf{b}}_{i_1}, \ldots, \underline{\mathbf{b}}_{i_p} - \underline{\mathbf{b}}_{i_1}\}$ *is linearly independent.*

The proof of this proposition is straightforward but tedious, and it is omitted in the interest of brevity. The proposition guarantees that when the columns of B are distinct, every strongly separated partition can be generated by Procedure PART-mD. When the points are not distinct, since PART-1D produces only two strongly separated partitions even when there are more, not every strongly separated partition will be generated by Procedure PART-mD. However, there will be at least one optimal partition in the set Φ, since for every strongly separated partition not in Φ there is one in Φ with the same cost. By using the fact that PART-1D generates only two partitions, it can be shown that the number of partitions $|\Phi|$ generated by Procedure PART-mD is no greater than $n^{p(p+1)/2}$ and that the computational complexity of Procedure PART-mD is $O(n^{p(p+3)/2})$. When $n_1 = n_2$ (i.e., for the GB problem), we have $|\Phi| \leqq \frac{1}{2}n^{p(p+1)/2}$.

These procedures for generating strongly separated partitions may be used to find optimal solutions for the GP problem on rank-one and other low-rank positive semidefinite matrices in polynomial time. This motivates seeking a low-rank approximation of a graph's connection matrix by the principal-components approach discussed in the following section.

5. The principal-components approximation. The heuristic approximation advocated by this paper is to seek a good low-rank approximation to the connection matrix B^tB of the GP problem by approximating the points $\underline{\mathbf{b}}_i$ by their projections on an appropriately chosen low-dimensional linear variety. The choice of variety is based on the following result attributed to Pearson [17]. To state the result we need to introduce a few notions. Suppose p, n, and k are given positive integers with $p \leqq \min \{n, k\}$. Also given is a $k \times n$ matrix B with columns $\underline{\mathbf{b}}_i \in \mathbb{R}^k$. Consider a pD linear variety $L(\underline{\mathbf{w}}, Z) \triangleq \{\underline{\mathbf{w}} + Z\underline{\mathbf{e}} : \underline{\mathbf{e}} \in \mathbb{R}^p\}$ in $\mathbb{R}^k$, parameterized by $\underline{\mathbf{w}} \in \mathbb{R}^k$ and a $k \times p$ matrix Z with orthonormal columns. For each $i \in \mathbf{V}$ let the *orthogonal projection* of $\underline{\mathbf{b}}_i$ onto $L(\underline{\mathbf{w}}, Z)$ be denoted by $P_{L(\underline{\mathbf{w}},Z)}(\underline{\mathbf{b}}_i) \triangleq \underline{\mathbf{w}} + ZZ^t(\underline{\mathbf{b}}_i - \underline{\mathbf{w}})$. It is the point on the variety $L(\underline{\mathbf{w}}, Z)$ that is *closest* to $\underline{\mathbf{b}}_i$ in the least-squares sense. Define a $k \times n$ matrix C as

$$C = B - \underline{\mathbf{m}}\underline{\mathbf{1}}^t; \qquad \underline{\mathbf{m}} \triangleq \frac{1}{n} \sum_{i=1}^{n} \underline{\mathbf{b}}_i. \tag{5.1}$$

Here $\underline{\mathbf{m}}$ is the *centroid* of the $\underline{\mathbf{b}}_i$'s, and the ith column of C is $\underline{\mathbf{c}}_i = \underline{\mathbf{b}}_i - \underline{\mathbf{m}}$ for each $i \in \mathbf{V}$. Let the singular-value decomposition (SVD) of C be given as

$$C = \sum_{i=1}^{\min\{k,n\}} \sigma_i \underline{\mathbf{u}}_i \underline{\mathbf{v}}_i^t = U\Sigma V^t, \tag{5.2}$$

where $\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_{\min(k,n)}$ are the singular values arranged in descending order, $\underline{\mathbf{u}}_i$ is the ith left singular vector, and $\underline{\mathbf{v}}_i$ is the ith right singular vector of C. We are now ready to state Pearson's result.

PEARSON'S PRINCIPAL-COMPONENTS THEOREM [4]. *For a given set of n points $\{\underline{\mathbf{b}}_i : i \in \mathbf{V}\}$ in $\mathbb{R}^k$, the pD linear variety $L(\underline{\mathbf{w}}, Z)$ that* minimizes *the total error* $\sum_{i=1}^{n} \|\underline{\mathbf{b}}_i - P_{L(\underline{\mathbf{w}},Z)}(\underline{\mathbf{b}}_i)\|^2$ is $L(\underline{\mathbf{m}}, \hat{U})$, *where $\underline{\mathbf{m}}$ is the centroid of these points as defined by* (5.1) *and the columns of $\hat{U} \triangleq [\underline{\mathbf{u}}_i, \underline{\mathbf{u}}_2, \ldots, \underline{\mathbf{u}}_p]$ are the left singular vectors corresponding to the p largest singular values of the matrix C of* (5.1) *and* (5.2). *Furthermore, if $\hat{\underline{\mathbf{b}}}_i \triangleq P_{L(\underline{\mathbf{m}},\hat{U})}(\underline{\mathbf{b}}_i)$ denotes the projection of $\underline{\mathbf{b}}_i$ on the variety $L(\underline{\mathbf{m}}, \hat{U})$, then the matrix $\hat{B}$ formed from these projections is given by*

$$\hat{B} = [\hat{\underline{\mathbf{b}}}_1, \hat{\underline{\mathbf{b}}}_2, \ldots, \hat{\underline{\mathbf{b}}}_n] = \underline{\mathbf{m}}\underline{\mathbf{1}}^t + \sum_{i=1}^{p} \sigma_i \underline{\mathbf{u}}_i \underline{\mathbf{v}}_i^t = \underline{\mathbf{m}}\underline{\mathbf{1}}^t + \hat{U}\hat{\Sigma}\hat{V}^t, \tag{5.3}$$

where $\hat{\Sigma} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$ *is the diagonal matrix with the p largest singular values of C and $\hat{V} = [\underline{\mathbf{v}}_1, \underline{\mathbf{v}}_2, \ldots, \underline{\mathbf{v}}_p]$ are the corresponding right singular vectors.*

For various interpretations of the principal-components approximation the reader is referred to an excellent tutorial on the subject by Rao [4].

The approximation suggested by Frankle and Karp to the multidimensional geometric clustering problem is a 1D approximation obtained by retaining only the components of the vectors along one of the coordinate axes. The axis chosen is the one on which the sum of squares of the components of the given vectors is maximum [7]. Unfortunately, the solution thus obtained is highly dependent on the choice of basis for the vectors or, equivalently, on the choice of the square root B of the connection matrix A. Frankle and Karp, therefore, specify which square root to use. The principal-components approximation, on the other hand, is insensitive to the choice of square root and does not constrain the search for the best variety to the set of coordinate axes. For this and other interpretations of the principal-components approximation of the set of vectors, the reader is referred to [5].

The principal-components heuristic for solving the GP problem on $A = B^tB$ is to solve the easier GP problem on p-rank matrix $\hat{B}^t\hat{B}$. The difference between the solutions to the two problems is quantified in § 8, where lower bounds on $\text{cost}_A(\underline{\mathbf{s}})$ are also presented. To solve the clustering problem on $\hat{B}$, given the connection matrix A, one need not explicitly determine a square root B of A. The following two propositions give a computationally efficient scheme for computing the entries of C^tC and $\hat{B}^t\hat{B}$ directly from the connection matrix A. Moreover, the discussion in § 6 shows that the clustering problem on $\hat{B}$ can, in fact, be solved by applying Procedure PART-mD to a subset of eigenvectors of C^tC.

Let I denote the $n \times n$ identity matrix and $J \triangleq \underline{1}\underline{1}^T$ denote the $n \times n$ matrix of ones. Define $P \triangleq I - (1/n)J$. It can be seen that P is the orthogonal projection matrix onto the orthogonal complement of $\underline{1}$.

PROPOSITION 5. *For a given connection matrix $A = B^tB$, consider the matrix C defined by* (5.1). *Then the matrix $M \triangleq C^tC$ can be constructed directly from A as $M = PAP$ in $O(n^2)$ time.*

Proof. From (5.1) we see that $C = B - (1/n)B\underline{1}\underline{1}^t = BP$. Therefore,

$$M \triangleq C^tC = PAP = A - \underline{1}\underline{\mathbf{r}}^t - \underline{\mathbf{r}}\underline{1}^t + \eta J, \tag{5.4}$$

where $\underline{\mathbf{r}} \triangleq (1/n)A\underline{1}$ is the *average row-sum* vector of A and $\eta \triangleq (1/n^2)\underline{1}^tA\underline{1}$ is the average of all entries in A. Note that $\underline{\mathbf{r}}$ and η can be computed from the entries of A in $O(n^2)$ time. Also, (5.4) indicates that the (i, j-th) entry of M can be obtained from the corresponding entry of A by subtracting r_i and r_j and then adding η. Hence computing M from A can be done in $O(n^2)$ time. This completes the proof. □

The matrix M defined by (5.4) has several interesting features. Note that $P\underline{1} = \underline{0}$ implies that $M\underline{1} = \underline{0}$. Moreover, let $\mu_1 \geqq \mu_2 \geqq \cdots \geqq \mu_n$ denote the eigenvalues of A arranged in descending order, and let $\lambda_1 \geqq \lambda_2 + \cdots \geqq \lambda_{n-1}$ denote $(n-1)$ eigenvalues of M that have eigenvectors orthogonal to $\underline{1}$. We will refer to these as the *nontrivial eigenvalues* of M. In addition to these, M always has a *trivial eigenvalue* of $\lambda_n = 0$ with $\underline{1}$ as its corresponding eigenvector. The following *interlacing relation* between the eigenvalues of A and the nontrivial eigenvalues of M can be established by using the *min-max principle* [18]:

$$\mu_i \geqq \lambda_i \geqq \mu_{i+1} \quad \text{for all} \quad 1 \leqq i \leqq n-1. \tag{5.5}$$

This shows that rank $(M) \leqq$ rank (A).

The singular values σ_i and right singular vectors $\underline{\mathbf{v}}_i$ of C needed in the principal-components approximation can be directly found from an eigendecomposition of M [19]. Since $M = C^tC$, the singular values of C are the square roots of the eigenvalues of M, and the right singular vectors of C are the eigenvectors of M. Thus the approximants $\hat{\Sigma}$ and $\hat{V}$ can be identified, respectively, with the square roots of the principal eigenvalues and the corresponding eigenvectors of M. Thus $\hat{V} = \hat{E} \triangleq [\underline{\mathbf{e}}_1, \underline{\mathbf{e}}_2, \ldots, \underline{\mathbf{e}}_p]$ and $\hat{\Sigma} = \hat{\Lambda}^{1/2} \triangleq \operatorname{diag}[\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_p^{1/2}]$.

We will see in the next section that an explicit determination of U is not needed to compute the solution to the GP problem on the approximation $\hat{A} = \hat{B}^t\hat{B}$. We will see that the matrix $\hat{B}$ also need not be constructed. Procedure PART-mD can be used directly on $\hat{E}$ to solve the lower-dimensional clustering problem. Thus only one eigendecomposition, that of M, is needed. We will go one step further. We will show that only a few eigenvalues and eigenvectors of M need to be computed.

6. The heuristic algorithms. In this section we list algorithms for solving the GP problem on rank-one and other low-rank principal-components-based approximations of the connection matrix. These algorithms use Procedures PART-1D and PART-mD of §§ 3 and 4. The basic idea underlying the algorithms listed in this section is to find a low-rank positive semidefinite approximation $\hat{A} = \hat{B}^t\hat{B}$ to a positive semidefinite connection matrix A by using the eigendecomposition of an intermediate matrix M and then to use Procedure PART-mD on $\hat{B}$. Among all the strongly separated partitions generated by this procedure, the one with the smallest value of $\text{cost}_A(\underline{\mathbf{s}})$ is reported as the output. Note that the low-rank approximation $\hat{A}$ is used only to generate a small set $\boldsymbol{\Phi}$ of candidate partitions, namely, those that are strongly separated with respect to $\hat{B}$. However, the partition from the set $\boldsymbol{\Phi}$ that is finally output is based on comparison of $\text{cost}_A(\underline{\mathbf{s}})$ and not of $\text{cost}_{\hat{A}}(\underline{\mathbf{s}})$. Thus the final output partition (say, $\underline{\mathbf{s}}_o$) *may not* be the solution to the GP problem on the approximated matrix $\hat{A}$, although the latter is also guaranteed to be in the set $\boldsymbol{\Phi}$.

Using the notation of § 5, recall that $\hat{C} = \hat{U}\hat{\Sigma}\hat{V}^t$ and that $\hat{B}$ defined in (5.3) is obtained by uniformly translating all the columns of $\hat{C}$ by the centroid vector $\underline{\mathbf{m}}$. However, the set of strongly separated partitions with respect to a collection of vectors is not affected by uniform translation of the vectors. Also, the set of strongly separated partitions is not affected by a rotation of the coordinate axes or by scaling of the coordinates of each vector identically. Thus the set of strongly separated partitions with respect to $\hat{B}$ is also the set of strongly separated partitions with respect to $\hat{V}^t = \hat{E}^t$. Therefore, Procedure PART-mD (or PART-1D, when $p = 1$) can be applied to $\hat{E}^t$ itself. Recall that the rows of $\hat{E}^t$ are the principal eigenvectors of matrix M.

It is instructive to note that although the set of strongly separated partitions is unaffected by operations such as uniform translation, the optimal partition itself can be affected. The following simple example serves to demonstrate this possibility.

Example 2. Consider a 1×3 matrix $\hat{C} = (-2\ -1\ 3)$ and $\hat{B} = (1\ 2\ 6)$ obtained by a uniform translation of $+3$. Here $n = 3$, and let $n_1 = 2$ and $n_2 = 1$. The set $\boldsymbol{\Phi}$ of strongly separated partitions with respect to both $\hat{B}$ and $\hat{C}$ is the same:

$$\boldsymbol{\Phi} = \{\underline{\mathbf{s}}_a, \underline{\mathbf{s}}_b\}, \quad \text{where} \quad \underline{\mathbf{s}}_a = (0\ 1\ 1)^t \quad \text{and} \quad \underline{\mathbf{s}}_b = (1\ 1\ 0)^t.$$

But $\text{cost}_{\hat{C}^t\hat{C}}(\underline{\mathbf{s}}_a) = -4$ is larger than $\text{cost}_{\hat{C}^t\hat{C}}(\underline{\mathbf{s}}_b) = -9$, whereas $\text{cost}_{\hat{B}^t\hat{B}}(\underline{\mathbf{s}}_a) = +8$ is smaller than $\text{cost}_{\hat{B}^t\hat{B}}(\underline{\mathbf{s}}_b) = +18$. Therefore, $\underline{\mathbf{s}}_b$ is optimal for GP on $\hat{C}^t\hat{C}$, whereas $\underline{\mathbf{s}}_a$ is optimal for GP on $\hat{B}^t\hat{B}$. Hence the optimal partition changed after translation, although the set $\boldsymbol{\Phi}$ itself was unaffected.

The example shows that even though the generation of $\boldsymbol{\Phi}$ can be based on the eigenvectors of M, the choice of partition from the set should be based on comparison of cost_A and not of cost_M. The algorithm based on 1D approximations is given below.

Algorithm GP-1D

```
**Input:** Two positive integers $n_1$ and $n_2$ such that $n_1 + n_2 = n$ and an $n \times n$ connection
matrix $A$.
**Output:** A feasible partition indicator vector $\underline{\mathbf{s}}_o$.
begin
(1) Obtain $M$ from $A$ by using (5.4);
(2) Find the largest eigenvalue and corresponding eigenvector $\underline{\mathbf{e}}_1$ of $M$;
(3) PART-1D ($\underline{\mathbf{e}}_1$, $n_1$, $n_2$, $\underline{\mathbf{s}}_a$, $\underline{\mathbf{s}}_b$);
(4) IF ($\text{cost}_A(\underline{\mathbf{s}}_a) \leqq \text{cost}_A(\underline{\mathbf{s}}_b)$) THEN $\underline{\mathbf{s}}_o = \underline{\mathbf{s}}_a$;
      ELSE $\underline{\mathbf{s}}_o = \underline{\mathbf{s}}_b$;
      ENDIF
end
```

Example 3. To illustrate that 1D approximations are not always enough to get optimal partitions even for problems with connection matrices of rank two, consider bisecting a six-node graph described by the connection matrix

$$A = \begin{bmatrix} 85 & 62 & 56 & 44 & 38 & 15 \\ 62 & 68 & 60 & 44 & 36 & 42 \\ 56 & 60 & 53 & 39 & 32 & 36 \\ 44 & 44 & 39 & 29 & 24 & 24 \\ 38 & 36 & 32 & 24 & 30 & 18 \\ 15 & 42 & 36 & 34 & 18 & 45 \end{bmatrix}.$$

Using (5.4), one can obtain from A the matrix M, which is rank two, whose eigenvalues are $\lambda_1 = 50$ and $\lambda_2 = 10$, and whose corresponding eigenvectors are

$$[\sqrt{\lambda_1}\underline{e}_1 \ \sqrt{\lambda_2}\underline{e}_2] = \begin{bmatrix} 5 & 0 \\ 0 & 2 \\ 0 & 1 \\ 0 & -1 \\ 0 & -2 \\ -5 & 0 \end{bmatrix}.$$

Procedure PART-1D on $\underline{e}_1$ will return one of six indicator vectors ($\underline{s}_1$ through $\underline{s}_6$) as $\underline{s}_a$ and will return $\underline{1} - \underline{s}_a$ as $\underline{s}_b$, depending on the strategy used to break ties in the middle four entries of $\underline{e}_1$.

$$\underline{s}_1 = (1\ 1\ 1\ 0\ 0\ 0)^t, \quad \underline{s}_6 = \underline{1} - \underline{s}_1, \quad \text{cost}_A(\underline{s}_1) = \text{cost}_A(\underline{s}_6) = 1416,$$

$$\underline{s}_2 = (1\ 1\ 0\ 1\ 0\ 0)^t, \quad \underline{s}_5 = \underline{1} - \underline{s}_2, \quad \text{cost}_A(\underline{s}_2) = \text{cost}_A(\underline{s}_5) = 1424,$$

$$\underline{s}_3 = (1\ 1\ 0\ 0\ 1\ 0)^t, \quad \underline{s}_4 = \underline{1} - \underline{s}_3, \quad \text{cost}_A(\underline{s}_3) = \text{cost}_A(\underline{s}_4) = 1425.$$

However, the six $\underline{s}_i$'s listed above do not all have the same cost with respect to M. It is clear in this case that the second principal component of M is needed for selecting the best bisection. The reader may verify that $\underline{s}_1$ is a strongly separated bisection with respect to C, whereas $\underline{s}_2$ and $\underline{s}_3$ are not even NICH bisections with respect to C. The algorithm based on mD approximations is listed below.

Algorithm GP-mD

```
    Input: Three positive integers $n_1$, $n_2$, and $p$, such that $n_1 + n_2 = n$, $2 \leqq p < n$, and
    an $n \times n$ connection matrix $A$.
    Output: A feasible partition $\underline{s}_o$.
    begin
    (1) Obtain $M$ from $A$ by using (5.4);
    (2) Find the $p$ largest eigenvalues of $M$ and corresponding eigenvectors $\hat{E}$ =
        $[\underline{e}_1, \underline{e}_2, \ldots, \underline{e}_p]$.
    (3) PART-mD ($\hat{E}^t$, $p$, $n_1$, $n_2$, $\Phi$);
    (4) low = $+\infty$;
    (5) FOR each $\underline{s} \in \Phi$ DO
        IF (cost$_A(\underline{s})$ < low) THEN
            $\underline{s}_o = \underline{s}$;
            low = cost$_A(\underline{s})$;
        END IF
        END FOR
    end
```

Since $\hat{M} = \hat{E}\hat{\Lambda}\hat{E}^t$ is only an approximation to M and $\hat{A}$ is only an approximation to A, the partition $\underline{s}_o$ obtained by GP-mD will not, in general, be an optimal solution to GP on A. Bounds on the proximity of the cost of $\underline{s}_o$ to the cost of the best partition are presented in § 8.

7. Choice of diagonal entries. Recall from § 2 that diagonal entries of A do not change the cost of a candidate partition with respect to A; hence the optimal solution to the GP problem on A remains unaffected. However, different diagonal choices for A indeed do affect the eigenvalues and eigenvectors of M, thereby altering the partition $\underline{s}_o$ generated by the principal-components heuristics of § 6. In this section we suggest two choices for the diagonal entries of the connection matrix A.

First observe from the A-to-M transformation of (5.4) that if A changes to $A + \varepsilon I$, then M changes to $M + \varepsilon P$ since P, being a projection matrix, is idempotent. Moreover, $(M + \varepsilon P)\underline{1} = \underline{0}$, and

$$M\underline{e} = \lambda\underline{e} \Rightarrow (M + \varepsilon P)\underline{e} = (\lambda + \varepsilon)\underline{e} \quad \text{for all} \quad \underline{e}^t\underline{1} = 0. \tag{7.1}$$

This means that 0 remains an eigenvalue of $M + \varepsilon P$ with eigenvector $\underline{1}$, whereas all other eigenvalues of M are translated by ε with the same set of eigenvectors. Therefore, the addition of εI to A does not change the partition generated by the principal-components heuristics of § 6.

However, nonconstant diagonals can affect the partition obtained and can even bring M closer to low rank. A similar problem of choosing diagonals arises in factor analysis in multivariate statistics, wherein cross correlations between multiple quantities are measured but autocorrelations (called communalities) have to be guessed [20]. The problem is one of determining a suitable diagonal for a symmetric matrix that makes the matrix positive semidefinite and close to low rank. Different choices of diagonal (such as the Spearman formula [21]) are suggested in the factor analysis literature, but our experiments with these choices indicate that the partitions of random graphs so obtained are not better than the partitions obtained with a constant diagonal.

One other interesting possibility that we suggested in [8] and have found to be useful is to choose the diagonal entries of A so as to make the new matrix have equal row sums and be positive semidefinite. This is done as follows. Assume that A has nonnegative entries and zeros on the diagonal initially. As before, let $\underline{r} = (1/n)A\underline{1}$ denote the average row-sum vector, and let $r_{\max} = \max_i (r_i)$ be the maximum average row sum. Then choose the ith diagonal entry of A as

$$a_{ii} = n(2r_{\max} - r_i). \tag{7.2}$$

This will make all the rows of the new A sum to $2nr_{\max}$. It can be easily verified that this new connection matrix A is simply $2nr_{\max}I - Q$, where Q is the Laplacian matrix of the graph. Hence the eigenvector for the second-smallest eigenvalue of the Laplacian is the same as the eigenvector for the second-largest eigenvalue of A, which is shown in § 8.2 to be equal to the eigenvector for the largest eigenvalue of M (see (8.20)). Therefore, Algorithm GP-1D applied after the diagonal is chosen according to (7.2) is essentially the same as the algorithm of Pothen, Simon, and Liou in [13].

8. Lower bounds. To evaluate the quality of the partitions obtained by the heuristics of § 6, we can compare their costs with lower bounds (on the cost of the optimal partition) established in this section. These bounds are easily computable from certain eigenvalues and eigenvectors of the matrix M used in Algorithm GP-mD. It should be intuitively obvious that the quality of the partition obtained by GP-mD depends intimately on the relative magnitude of the $n - p$ eigenvalues of M dropped by the principal-components approximation. In this section we will quantify this notion. We will see that for each choice of rank $p \geqq 1$ we can obtain a lower bound LB_p on the optimal cost. Furthermore, increasing the value of p improves the lower bound LB_p. Finally, for the special case of bisectioning we will show that our lower bounds are better than the lower bound established by Donath and Hoffman [22], which is computed directly from the two largest eigenvalues of A. The same conclusions will be reached in the more general GP problem ($n_1 \neq n_2$) if the connection matrix A has equal row sums.

For deriving the lower bounds of this section we will assume that $M = PAP$ has eigenvalues $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_{n-1} > 0 = \lambda_n$, i.e., that M is positive semidefinite and that the trivial zero eigenvalue of M is simple. By using (7.1) this can be guaranteed by simply adding a sufficiently large constant diagonal to A. We will later see that to calculate the bounds in practice not all nontrivial eigenvalues of M need be positive. Also, for each $i \in \mathbf{V}$ let $\underline{\mathbf{e}}_i$ denote the normalized eigenvector of M corresponding to λ_i. Note that $\underline{\mathbf{e}}_n = (1/\sqrt{n})\underline{\mathbf{1}}$ and that $E = [\underline{\mathbf{e}}_1, \underline{\mathbf{e}}_2, \ldots, \underline{\mathbf{e}}_n]$ is an orthonormal matrix.

Without loss of generality, assume $n_1 \geqq n_2$ throughout this section, i.e., n_1 denotes the larger part size. Consider any feasible partition indicated by $\underline{\mathbf{s}} \in \Gamma$. The key to deriving our lower bounds is to decompose $\mathrm{cost}_A(\underline{\mathbf{s}})$ into contributions from the separate eigenvectors of M as follows. Using (5.4) and denoting the row-sum averages of A by $\underline{\mathbf{r}} = (1/n)A\underline{\mathbf{1}}$ and the average of all entries of A by $\eta = (1/n^2)\underline{\mathbf{1}}^t A \underline{\mathbf{1}}$, we see that

$$\mathrm{cost}_A(\underline{\mathbf{s}}) = n_1^2\eta - (n_1 - n_2)\underline{\mathbf{s}}^t\underline{\mathbf{r}} - \sum_{i=1}^{n} \lambda_i(\underline{\mathbf{s}}^t\underline{\mathbf{e}}_i)^2, \tag{8.1}$$

where λ_i is the ith-largest eigenvalue of M and $\underline{\mathbf{e}}_i$ is the corresponding normalized eigenvector. The first term in the above cost decomposition is a constant independent of $\underline{\mathbf{s}} \in \Gamma$. We now attempt to upper bound the second and third terms by quantities that are independent of $\underline{\mathbf{s}} \in \Gamma$, which thereby results in a lower bound for $\mathrm{cost}_A(\underline{\mathbf{s}})$.

Sort the entries of the average row-sum vector $\underline{\mathbf{r}}$, and let r^* denote the sum of the n_1 *largest* entries of $\underline{\mathbf{r}}$. Clearly, this implies

$$(n_1 - n_2)\underline{\mathbf{s}}^t\underline{\mathbf{r}} \leqq (n_1 - n_2)r^* \quad \text{for all} \quad \underline{\mathbf{s}} \in \Gamma \tag{8.2}$$

since $n_1 \geqq n_2$ by assumption. Note that (8.2) provides the desired upper bound for the second term of (8.1).

Now consider the third term $\sum_{i=1}^{n-1} \lambda_i(\underline{\mathbf{s}}^t\underline{\mathbf{e}}_i)^2$ in (8.1) for any $\underline{\mathbf{s}} \in \Gamma$. Note that the index i need range only from 1 to $n-1$ since $\lambda_n = 0$. In order to upper-bound this term we start by listing some constraints to be satisfied by the quantities $(\underline{\mathbf{s}}^t\underline{\mathbf{e}}_i)^2$. Using $EE^t = I$ and $\underline{\mathbf{s}}^t\underline{\mathbf{s}} = n_1$, we have $\sum_{i=1}^{n} (\underline{\mathbf{s}}^t\underline{\mathbf{e}}_i)^2 = n_1$. Moreover, $\underline{\mathbf{e}}_n = (1/\sqrt{n})\underline{\mathbf{1}}$; hence $(\underline{\mathbf{s}}^t\underline{\mathbf{e}}_n)^2 = n_1^2/n$. Therefore,

$$\sum_{i=1}^{n-1} (\underline{\mathbf{s}}^t\underline{\mathbf{e}}_i)^2 = \frac{n_1 n_2}{n} \quad \text{for all} \quad \underline{\mathbf{s}} \in \Gamma. \tag{8.3}$$

For each $i = 1, 2, \ldots, n-1$ run PART-1D ($\underline{\mathbf{e}}_i, n_1, n_2, \underline{\mathbf{s}}_{a,i}, \underline{\mathbf{s}}_{b,i}$) and define

$$\beta_i \triangleq \max\{(\underline{\mathbf{e}}_i^t\underline{\mathbf{s}}_{a,i})^2, (\underline{\mathbf{e}}_i^t\underline{\mathbf{s}}_{b,i})^2\}. \tag{8.4}$$

By using (3.1) it follows that

$$0 \leqq (\underline{\mathbf{s}}^t\underline{\mathbf{e}}_i)^2 \leqq \beta_i \quad \text{for all} \quad \underline{\mathbf{s}} \in \Gamma. \tag{8.5}$$

Consider a fixed integer $p \geqq 1$ but less than n. Define $\hat{E} = [\underline{\mathbf{e}}_1, \underline{\mathbf{e}}_2, \ldots, \underline{\mathbf{e}}_p]$ and $\hat{\Lambda}^{1/2} \triangleq \mathrm{diag}\,[\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_p^{1/2}]$, and obtain the set Φ of strongly separated partitions with respect to $\hat{E}^t$ by using either PART-1D or PART-mD. Note that the set Φ is automatically obtained in step (3) of the principal-components heuristics GP-mD (or GP-1D if $p = 1$) applied on A for the chosen value of p. Since the set Φ is small (no greater than $n^{p(p+1)/2}$), the following maximum may be computed relatively easily:

$$d \triangleq \max_{\underline{s} \in \Phi} \sum_{i=1}^{p} \lambda_i (\underline{s}^t \underline{e}_i)^2. \tag{8.6}$$

If $\hat{M} = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i^t = \hat{E} \hat{\Lambda} \hat{E}^t$, then $\text{cost}_{\hat{M}}(\underline{s}) = -\underline{s}^t \hat{M} \underline{s} = -\sum_{i=1}^{p} \lambda_i (\underline{s}^t \underline{e}_i)^2$. By Proposition 2 an optimal partition for GP on $\hat{M}$ must be strongly separated with respect to $\hat{\Lambda}^{1/2} \hat{E}^t$. Consequently,

$$\text{for all} \quad \underline{s} \in \Gamma, \qquad \sum_{i=1}^{p} \lambda_i (\underline{s}^t \underline{e}_i)^2 \leqq \max_{\underline{s} \in \Phi} \sum_{i=1}^{p} \lambda_i (\underline{s}^t \underline{e}_i)^2 = d. \tag{8.7}$$

Define the following subset of $\mathbb{R}^{n-1}$:

$$\Psi \triangleq \left\{ \underline{x} \in \mathbb{R}^{n-1} : 0 \leqq x_i < \beta_i, \ \sum_{i=1}^{n-1} x_i = \frac{n_1 n_2}{n} \text{ and } \sum_{i=1}^{p} \lambda_i x_i \leqq d \right\}, \tag{8.8}$$

where the β_i's and d are defined by (8.4) and (8.6), respectively. It follows from the constraints on $(\underline{s}^t \underline{e}_i)^2$ in (8.3), (8.5), and (8.7) that

$$\max_{\underline{s} \in \Gamma} \sum_{i=1}^{n-1} \lambda_i (\underline{s}^t \underline{e}_i)^2 \leqq \max_{\underline{x} \in \Psi} \sum_{i=1}^{n-1} \lambda_i x_i. \tag{8.9}$$

The following proposition indicates that the right-hand side of (8.9) can be computed easily from β_i, d, and the eigenvalues λ_i. It can then be used to bound the third term in the cost expression of (8.1).

PROPOSITION 6. *Let the eigenvalues of M be $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_{n-1} > 0 = \lambda_n$. Let $p \geqq 1$ be a fixed integer smaller than n. Define β_i as in* (8.4), *and define d as in* (8.6). *Let $k \geqq 1$ be the smallest integer such that*

$$\sum_{i=1}^{k} \lambda_i \beta_i \geqq d. \tag{8.10}$$

Similarly, let $l \geqq p + 1$ be the smallest integer such that

$$\sum_{i=1}^{k-1} \beta_i + \frac{1}{\lambda_k} \left[d - \sum_{i=1}^{k-1} \lambda_i \beta_i \right] + \sum_{i=p+1}^{l} \beta_i \geqq \frac{n_1 n_2}{n}. \tag{8.11}$$

Then the vector $\hat{\underline{x}} \in \mathbb{R}^{n-1}$ with entries defined by

$$\hat{x}_i \triangleq \begin{cases} \beta_i & \text{for } 1 \leqq i \leqq k-1, \\ \dfrac{d}{\lambda_k} - \displaystyle\sum_{i=1}^{k-1} \beta_i \frac{\lambda_i}{\lambda_k} & \text{for } i = k, \\ 0 & \text{for } k+1 \leqq i \leqq p, \\ \beta_i & \text{for } p+1 \leqq i \leqq l-1, \\ \dfrac{n_1 n_2}{n} - \displaystyle\sum_{i=1}^{l-1} \hat{x}_i & \text{for } i = l, \\ 0 & \text{for } l+1 \leqq i \leqq n-1 \end{cases} \tag{8.12}$$

achieves the maximum of $\sum_{i=1}^{n-1} \lambda_i x_i$ over all $\underline{x}$ from the polytope Ψ defined by (8.8).

The proof of this proposition is not included in the interest of brevity. It follows from standard results in linear programming. The above proposition can be used to prove the following main result, which provides a lower bound for the cost of an optimal partition of GP on a positive-definite connection matrix A.

PROPOSITION 7. *Let A be an $n \times n$ positive-definite connection matrix, and let $p \geqq 1$ be an integer less than n. Also, let $n_1 \geqq n_2 \geqq 1$ be two integers such that $n_1 + n_2 = n$. Then for all $\underline{s} \in \Gamma$, $\operatorname{cost}_A(\underline{s})$ as defined in* (1.1) *is bounded as*

$$\operatorname{cost}_A(\mathbf{s}) \geqq \mathrm{LB}_p \triangleq n_1^2\eta - (n_1 - n_2)r^* - \sum_{i=1}^{l} \lambda_i \hat{x}_i, \tag{8.13}$$

where $\eta = (1/n^2)\underline{1}'A\underline{1}$ is the average of all entries of A, r^* denotes the sum of the n_1 largest entries of $\underline{r} = (1/n)A\underline{1}$, the λ_i's are the eigenvalues of $M = PAP$ in descending order, and $\hat{\underline{x}}$ is defined by (8.12).

Proof. From the interlacing property (5.5) between the eigenvalues of A and M, we note that if A is positive definite, then M is positive semidefinite and the zero eigenvalue is simple, i.e., the eigenvalues of M satisfy $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_{n-1} > 0 = \lambda_n$. From (8.1) and (8.2) and by using $\lambda_n = 0$, we see that

$$\text{for all} \quad \underline{s} \in \Gamma, \qquad \operatorname{cost}_A(\underline{s}) \geqq n_1^2\eta - (n_1 - n_2)r^* - \sum_{i=1}^{n-1} \lambda_i(\underline{s}'\underline{e}_i)^2.$$

From (8.9) and the definitions of l and $\hat{x}$ in Proposition 6 we can upper-bound the last term in the right-hand side of the above inequality by $\sum_{i=1}^{l} \lambda_i \hat{x}_i$. This completes the proof. □

Note that to compute the lower bound LB_p all the eigenvalues and eigenvectors of M need not be computed. It suffices to compute only the l largest eigenvalues and the corresponding eigenvectors, where l is defined by (8.11) in Proposition 6. If in a specific instance it turns out that l is large and the computation of the l largest eigenvalues and corresponding eigenvectors of M is prohibitively expensive, then the following weaker lower bound $\tilde{\mathrm{LB}}_p$ may be computed by using only the $\tilde{l}$ largest eigenvalues of M, for any $\tilde{l}$ such that $p < \tilde{l} < l$.

$$\tilde{\mathrm{LB}}_p \triangleq n_1^2\eta - (n_1 - n_2)r^* - \sum_{i=1}^{\tilde{l}} \lambda_i \tilde{x}_i$$

where

$$\tilde{x}_i \triangleq \begin{cases} \hat{x}_i & \text{for } 1 \leqq i \leqq \tilde{l} - 1, \\ \dfrac{n_1 n_2}{n} - \displaystyle\sum_{i=1}^{\tilde{l}-1} \hat{x}_i & \text{for } i = \tilde{l}, \\ 0 & \text{for } \tilde{l} + 1 \leqq i \leqq n - 1. \end{cases}$$

It is easily verified that $\mathrm{LB}_p \geqq \tilde{\mathrm{LB}}_p$ and that $\mathrm{LB}_{p+1} \geqq \mathrm{LB}_p$ and $\tilde{\mathrm{LB}}_{p+1} \geqq \tilde{\mathrm{LB}}_p$.

In the derivation of the lower bound, the positivity of all nontrivial eigenvalues of M was never actually needed. For the bounds to be valid, it suffices to have only the $\tilde{l}$ largest eigenvalues of M be positive. When rank-one approximations are used ($p = 1$), the expressions for lower bounds are considerably simpler because $d = \lambda_1\beta_1$, k as defined in Proposition 6 becomes 1, and l defined in Proposition 6 is the smallest integer greater than unity that satisfies

$$\sum_{i=1}^{l} \beta_i \geqq \frac{n_1 n_2}{n}.$$

For comparison purposes we include below a lower bound due to Donath and Hoffman [22].

THEOREM OF DONATH AND HOFFMAN. *If A is any $n \times n$ connection matrix and $\underline{\mathbf{s}} \in \Gamma$ is any indicator vector of a feasible partition with $n_1 \geqq n_2$, then*

$$\text{cost}_A(\underline{\mathbf{s}}) \geqq \text{LB}_{\text{DH}} \triangleq \tfrac{1}{2}(\underline{\mathbf{1}}^T A \underline{\mathbf{1}} - n_1\mu_1 - n_2\mu_2), \tag{8.14}$$

where μ_1 and μ_2 are the largest and second-largest eigenvalues of A, respectively.

Note that the bound LB_{DH} is computed directly from the eigenvalues of the connection matrix A, whereas LB_p and $\tilde{\text{LB}}_p$ for all values of p are computed by using eigenvalues of $M = PAP$. We now consider some special cases.

8.1. Graph bisectioning. In the case of bisectioning, $n_1 = n_2 = n/2$. Therefore, the lower bound of (8.13) reduces to

$$\text{LB}_p = \tfrac{1}{4}\underline{\mathbf{1}}^t A \underline{\mathbf{1}} - \sum_{i=1}^{l} \lambda_i \hat{x}_i,$$

where k is defined in (8.10) as before but $l \geqq p+1$ is the smallest integer satisfying

$$\frac{1}{\lambda_k}\left[d - \sum_{i=1}^{k-1} \lambda_i \beta_i\right] + \sum_{i=1}^{k-1} \beta_i + \sum_{i=p+1}^{l} \beta_i \geqq \frac{n}{4}.$$

The entries of $\hat{\underline{\mathbf{x}}}$ are defined by (8.12), as before. Similar results can be obtained for the bound $\tilde{\text{LB}}_p$. Note that the average row-sum vector $\underline{\mathbf{r}}$ does not contribute to the bounds in this special case.

PROPOSITION 8. *For the graph bisectioning problem*, $\text{LB}_1 \geqq \text{LB}_{\text{DH}}$.

The proof of this result is included in the appendix.

8.2. Equal-row-sum connection matrix. Recall that $\underline{\mathbf{1}}$ is an eigenvector of $M = PAP$ for the trivial eigenvalue $\lambda_n = 0$. If $\underline{\mathbf{1}}$ is also an eigenvector of A, i.e., if $A\underline{\mathbf{1}} = \alpha\underline{\mathbf{1}}$, then the connection matrix A has all row sums equal to α. Since A has only nonnegative entries, α must equal the largest eigenvalue of A. Since A is symmetric and $\underline{\mathbf{1}}$ is an eigenvector of A, all other eigenvalues of A must have eigenvectors orthogonal to $\underline{\mathbf{1}}$. Suppose $\underline{\mathbf{e}}$ is one such eigenvector of A corresponding to eigenvalue μ, i.e., $A\underline{\mathbf{e}} = \mu\underline{\mathbf{e}}$ and $\underline{\mathbf{e}}^t\underline{\mathbf{1}} = 0$. Since $P\underline{\mathbf{e}} = \underline{\mathbf{e}}$ in this case, we have $M\underline{\mathbf{e}} = PAP\underline{\mathbf{e}} = PA\underline{\mathbf{e}} = \mu P\underline{\mathbf{e}} = \mu\underline{\mathbf{e}}$, which implies that $\underline{\mathbf{e}}$ is also an eigenvector of M with the same eigenvalue μ. Therefore, if the eigenvalues of A are $\alpha = \mu_1 \geqq \mu_2 \geqq \cdots \geqq \mu_n$, then the nontrivial eigenvalues of M are

$$\lambda_i = \mu_{i+1} \quad \text{for all} \quad 1 \leqq i \leqq n-1 \tag{8.15}$$

in this special case.

PROPOSITION 9. *If A has equal row sums, then* $\text{LB}_1 \geqq \text{LB}_{\text{DH}}$.

The proof of this result is also found in the appendix.

The term $(n_1 - n_2)\underline{\mathbf{s}}^t\underline{\mathbf{r}}$ in the cost expression of (8.1) contributes negatively as $(n_1 - n_2)r^*$ to the lower bound of (8.13). The presence of this term could make the bound loose, especially because it is not weighted down by the small eigenvalues of M. This term may be large even when λ_{p+1} and the other eigenvalues dropped out in the principal-components approximation are small. When these eigenvalues are small, the principal-components heuristics provide good partitions but the lower bounds may be poor. The effect of this term on the tightness of the lower bound depends on the variation in the term over the set of all feasible partitions. In the bisectioning problem this term is zero and thus has no effect on the lower bound. For unequal-size partitions, if the connection matrix A has equal row sums, then this term is constant and does not affect the tightness of the lower bound. In both these special cases we have shown that the lower bound of (8.13) is superior to the Donath–Hoffman lower bounds. In the general case of unequal-size partitions and unequal row sums of A, we have not been able to prove that the lower bounds of (8.13) are superior to the Donath–Hoffman bound. However, the extensive simulation results of § 9 seem to suggest that this is true.

The tightness of the lower bounds of (8.13), in general, depends on the magnitude of $(n_1 - n_2)(r^* - r_*)$, where $r_* \triangleq \min_{\underline{s} \in \Gamma} \underline{s}^t \underline{r}$. The smaller the differences $n_1 - n_2$ or $r^* - r_*$, the tighter is the lower bound of (8.13).

8.3. Improved 1D bounds. The lower bound LB_1 can be further improved by using the second-best solution to the GP problem on $\hat{M} = \lambda_1 \underline{e}_1 \underline{e}_1^t$, as was first suggested in [5]. Consider, once again, the β_i's defined by (8.4) and $\Phi = \{\underline{s}_a, \underline{s}_b\}$ output by Procedure PART-1D ($\underline{e}_1$, n_1, n_2, $\underline{s}_a$, $\underline{s}_b$). Let δ_a be the difference between the n_1th-largest and $(n_1 + 1)$th-largest entries in $\underline{e}_1$, and let δ_b be the difference between the n_2th-largest and $(n_2 + 1)$th-largest entries in $\underline{e}_1$. They can be obtained by two additional calls to procedure SELECT. Also, define $\Gamma_1 \triangleq \Gamma - \Phi$ as the set of all feasible partitions other than $\underline{s}_a$ and $\underline{s}_b$, and observe that

$$\underline{e}_1^t \underline{s}_b + \delta_b \leqq \underline{e}_1^t \underline{s} \leqq \underline{e}_1^t \underline{s}_a - \delta_a \quad \text{for all} \quad \underline{s} \in \Gamma_1.$$

Therefore,

$$(8.16) \quad 0 \leqq (\underline{s}^t \underline{e}_1)^2 \leqq \beta_1^\# \triangleq \max \{(\underline{e}_1^t \underline{s}_a - \delta_a)^2, \underline{e}_1^t \underline{s}_b + \delta_b)^2\} \quad \text{for all} \quad \underline{s} \in \Gamma_1.$$

It is apparent that $\beta_1^\# \leqq \beta_1$. When $\beta_1^\#$ is strictly smaller than β_1, an improved lower bound can be obtained for $\mathrm{cost}_A(\underline{s})$, $\underline{s} \in \Gamma_1$, by replacing the constraint for $i = 1$ in (8.5) by (8.16) and by repeating the arguments of Propositions 6 and 7. The exercise leads to

$$\mathrm{cost}_A(\underline{s}) \geqq \mathrm{LB}_1^\# \triangleq n_1^2 \eta - (n_1 - n_2) r^* - \sum_{i=1}^{l} \lambda_i x_i^\# \quad \text{for all} \quad \underline{s} \in \Gamma_1,$$

where $l \geqq 2$ is the smallest integer that satisfies

$$\beta_1^\# + \sum_{i=2}^{l} \beta_i \geqq \frac{n_1 n_2}{n}$$

and the vector $\underline{x}^\#$ has entries defined as

$$x_i^\# \triangleq \begin{cases} \beta_1^\# & \text{for } i = 1, \\ \beta_i & \text{for } 2 \leqq i \leqq l - 1, \\ \dfrac{n_1 n_2}{n} - \beta_1^\# - \displaystyle\sum_{j=2}^{l-1} \beta_j & \text{for } i = l, \\ 0 & \text{for } l + 1 \leqq i \leqq n - 1, \end{cases}$$

Let $\underline{s}_o$ denote the feasible partition produced by Algorithm GP-1D on A. Obviously, $\underline{s}_o \in \Phi$. Therefore, $\mathrm{cost}_A(\underline{s}_o) \leqq \mathrm{LB}_1^\#$ is a sufficient condition for the optimality of $\underline{s}_o$. If $\mathrm{cost}_A(\underline{s}_o) > \mathrm{LB}_1^\#$, then $\mathrm{LB}_1^\#$ will serve as a lower bound on $\mathrm{cost}_A(\underline{s})$ for all $\underline{s} \in \Gamma$. This is a tighter bound on $\mathrm{cost}_A(\underline{s})$ over $\underline{s} \in \Gamma$ since $\beta_1^\# \leqq \beta_1$ implies that $\mathrm{LB}_1^\# \geqq \mathrm{LB}_1$. Furthermore, when $\beta_1^\#$ is strictly smaller than β_1, $\mathrm{LB}_1^\#$ is strictly greater than LB_1. As before, when only $\tilde{l} < l$ eigenvalues and eigenvectors of M have been computed, a weaker bound can be similarly found.

9. Experimental results. The algorithms described in the previous sections have been extensively tested on a large number of 0–1 connection matrices drawn from the following three categories:

(1) Sparse matrices (0–1 pattern only) from the Boeing–Harwell collection [23].

(2) Connection matrices of five-point and nine-point grid graphs [13].

(3) Connection matrices generated randomly.

In each case all the nondiagonal entries in the connection matrix were either zero or one. For results of simulations on graphs with more general, nonunit weights on edges, the reader is referred to [5], [8]–[10]. Sparsity control was included in the random matrix generator to control the number of nonzero off-diagonal entries in the matrices. We suggest the following indicator of sparsity, to be called the *density factor*:

$$\gamma(A) \triangleq \frac{\underline{\mathbf{1}}^t A \underline{\mathbf{1}} - \text{Trace}\,(A)}{n(n-1)a_{\max}}$$

where $a_{\max}$ is the largest nondiagonal entry in the $n \times n$ connection matrix A. When all nonzero entries in the off-diagonal positions are 1, as in the simulations reported here, γ is simply the fraction of nondiagonal entries that are not zero. The algorithms implemented were the following, based on Algorithms GP-1D and GP-mD of § 6:

GP-1D.0 is GP-1D applied on the connection matrix A with zero diagonal entries.

GP-1D.E is GP-1D applied on the A with diagonal entries chosen as per (7.2).

GP-2D.0 is GP-mD with $p = 2$ applied on A with zero diagonal entries.

GP-2D.E is GP-mD with $p = 2$ applied on A with diagonal entries chosen as per (7.2).

Each of the above algorithms generates a feasible partition as well as a lower bound on the cost of the optimal partition. The integer ceiling of the computed lower bounds is reported here since A has only integer entries and costs are hence integer valued. In most cases the computation of $\text{LB}_1^{\#}$ required at most three largest eigenvalues and eigenvectors of M, i.e., $l \leqq 3$, whereas LB_2 required $l \leqq 4$. The cases in which better bounds are possible with more eigenvalues are indicated by an ampersand (&) in the following tables. It must be mentioned, once again, that Algorithm GP-1D.E is equivalent to the algorithm of [13].

The above four algorithms were coded in FORTRAN and used special subroutines, provided by the standard IMSL package, to compute eigenvalues and eigenvectors of matrices. All experimental simulations were performed on a SUN-4 Sparc-station.

We begin by presenting our results on the bisectioning of some graphs with connection matrices from the Boeing–Harwell collection and some grid graphs in Table 1. In each case, $n_1 = n/2$ if n is even and $n_1 = (n + 1)/2$ if n is odd. The results of Table 1 indicate that Algorithm GP-2D.E gave the best performance in terms of both the cost and the lower bound, with the exception of BCSSTK04, in which GP-2D.0 outperformed the others. Also, the 2D algorithms outperformed their 1D counterparts in all cases, as expected.

The graphs in Table 1 are extremely sparse with $\gamma < 0.1$, with the single exception of BCSSTK04, which has $\gamma = 0.203$. In the case of BCSSTK04, the zero-diagonal Algorithms GP-1D.0 and GP-2D.0 gave significantly better lower bounds than did their equal-row-sum counterparts. Moreover, GP-2D.0 gave the best partition in this case. The results in the next table indicate that for graphs with higher density, the zero diagonal is a better choice than the equal-row-sum diagonal.

In Table 2 we present results on bisectioning randomly generated graphs with a density factor $\gamma = 0.3$. The zero-diagonal Algorithms GP-1D.0 and GP-2D.0 performed consistently better than their equal-row-sum counterparts. For example, in the $n = 500$ case the GP-2D.0 algorithm was able to produce a lower bound within 4% of the cost of the bisection generated. The performance of all four algorithms improves as n increases.

Table 3 presents the results for unequal-size partitioning on some randomly generated

TABLE 1
Bisectioning of Boeing–Harwell and grid graphs.

		GP-1D.0		GP-1D.E		GP-2D.0		GP-2D.E	
Name	n	cost	$\mathrm{LB}_1^\#$	cost	$\mathrm{LB}_1^\#$	cost	LB_2	cost	LB_2
NOS4	100	14	−6	14	3	14	−3	14	5
BCSSTK04	132	308	211	304	157	302	215	304	159
BCSSTK22	138	12	−45	14	0&	11	−39	4	1
DWT19393	193	142	−5	142	54	142	−2	139	56
DWT20909	209	97	−43	67	10	87	−35	50	13
ASH292	292	101	−57	23	4	54	−50	21	4
DWT31010	310	26	−44	20	3	20	−40	20	4
PLAT362	362	126	−65	125	23	122	−64	121	24
LSHP4066	406	41	−14	45	5	41	−13	40	5
NOS5	468	201	−22	212	50	194	4	182	62
GRD5.10.20	200	10	−5	10	2	10	−4	10	2
GRD9.10.20	200	28	−15	28	6	28	−12	28	6
GRD5.19.25	475	21	−11	20	3	20	−9	20	3
GRD9.19.25	475	57	−32	56	7	56	−30	56	7

graphs of density $\gamma = 0.3$ as well as on some Boeing–Harwell and grid graphs. In each case the size of the larger part n_1 was chosen to be approximately $0.6n$. For the GP-1D.0 case we also list the Donath–Hoffmann lower bound $\mathrm{LB}_{\mathrm{DH}}$ for comparison.

For random graphs with density of $\gamma = 0.3$ the performance of algorithms GP-1D.E and GP-2D.E was generally worse than their zero-diagonal counterparts in terms of both the cost of the partition generated and the computed lower bound. Finally, we present some statistics on the performance of Algorithms GP-1D.0 and GP-2D.0 as a function of the graph size n and the density factor γ. In this study we considered only the bisectioning case. We also omitted GP-1D.E and GP-2D.E from this study since the earlier discussed results indicate that the performance of these algorithms is not as good as their zero-diagonal counterparts for larger densities. The performance measure for each algorithm was chosen to be

$$\Delta = \left[1 - \frac{\mathrm{LB}}{\mathrm{cost}}\right] \times 100.$$

TABLE 2
Bisectioning on random graphs with $\gamma = 0.3$.

		GP-1D.0		GP-1D.E		GP-2D.0		GP-2D.E	
Name	n	cost	$\mathrm{LB}_1^\#$	cost	$\mathrm{LB}_1^\#$	cost	LB_2	cost	LB_2
RAND30.3	30	53	41	51	33	47	44	50	34
RAND50.3	50	150	126	172	103	145	126	152	106
RAND100.3	100	615	545	652	473	607	547	640	478
RAND150.3	150	1405	1275	1502	1084&	1395	1280	1470	1132&
RAND200.3	200	2587	2379	2719	2110&	2559	2383	2661	2144&
RAND300.3	300	6059	5629	6327	5068&	5976	5639	6197	5123&
RAND400.3	400	10829	10189	11334	9203&	10762	10225	11238	9234&
RAND500.3	500	17229	16531	18036	14963&	17146	16480	17731	15037&

TABLE 3
Unequal-size partitioning.

			GP-1D.0			GP-1D.E		GP-2D.0		GP-2D.E	
Name	n	n_1	cost	$LB_1^\#$	LB_{DH}	cost	$LB_1^\#$	cost	LB_2	cost	LB_2
RAND30.3	30	18	46	33	27	48	31	45	34	47	32
RAND50.3	50	30	141	109	90	156	98&	135	109	143	100
RAND100.3	100	60	590	488	412	618	452&	580	490	610	459
RAND200.3	200	120	2487	2193	1861	2579	2022&	2464	2197	2531	2054&
NOS4	100	60	15	−16	−21	14	3	14	−14	14	5
BCSPWR03	118	71	12	−33	−59	4	2	5	−32	4	2
BCSSTK05	153	92	174	−83	−218	70	20	62	−64	66	26
GRD5.10.20	200	120	22	−12	−17	10	3	10	−11	10	3
GRD9.10.20	200	120	50	−35	−49	28	6	28	−32	28	6

It is a measure of relative separation between cost achieved and the lower bound computed. Note that a small value of Δ indicates good performance of the algorithm in terms of both the cost of the partition generated and the lower bound. A measure of relative separation (such as Δ) is preferred over a measure of absolute separation (such as cost − LB) because scaling of the connection matrix affects absolute separation but does not change the relative separation. The results for graphs of fixed size $n = 100$ and density factor γ varying between 0.1 and 0.9 are shown in Table 4. In this table the mean Δ_{mean} and standard deviation Δ_{sd} of the performance measure were obtained by running both GP-1D.0 and GP-2D.0 on 11 random graphs for a fixed value of γ. These results show that GP-2D.0 performs better than does GP-1D.0 on average at all levels of density. Moreover, the performance of either algorithm improves dramatically as the density factor increases.

The final set of results for graphs with a fixed density of $\gamma = 0.3$ and size n varying between 30 and 500 are shown in Table 5. In this table the values of Δ_{mean} and Δ_{sd} were obtained by running both GP-1D.0 and GP-2D.0 on 10 random graphs for each value of n. The performance of either algorithm, once again, improves quite significantly on average as n increases. For example, on a random graph on 500 vertices the GP-2D.0 algorithm is expected to give a lower bound within 5% of the cost of the bisection generated, whereas on a graph on 100 vertices the lower bound is expected to be within 10% of the

TABLE 4
Statistics of performance versus density.

	GP-1D.0		GP-2D.0	
γ	Δ_{mean}	Δ_{sd}	Δ_{mean}	Δ_{sd}
0.1	32.2552	4.51599	24.8996	3.23034
0.2	16.2231	1.93971	11.8037	1.73474
0.3	11.5375	0.94880	8.0821	1.15707
0.4	9.1858	0.89699	6.4410	0.96194
0.5	7.0142	0.64616	4.8301	0.41526
0.6	5.5250	0.53461	3.9802	0.25564
0.7	4.5473	0.35456	3.2629	0.34936
0.8	3.4856	0.31932	2.6135	0.24535
0.9	2.4210	0.28736	1.8271	0.19292

TABLE 5
Statistics of performance versus graph size.

	GP-1D.0		GP-2D.0	
n	Δ_{mean}	Δ_{sd}	Δ_{mean}	Δ_{sd}
30	18.1527	4.16481	8.96826	3.62944
50	16.7834	1.68350	9.57364	1.62944
100	11.3557	0.79174	8.07255	1.21314
150	9.0417	0.55910	7.32810	0.58846
200	8.0465	0.89136	6.36207	0.45827
300	6.9040	0.36230	5.71936	0.22667
400	5.9739	0.43197	5.09929	0.38692
500	5.2265	0.20352	4.62660	0.18563

cost of the bisection generated. The results of Tables 4 and 5 are only for bisectioning; for unequal-size partitioning the values of Δ_{mean} will be larger as the difference between the sizes of the two parts increases.

In conclusion, this paper deals with a new class of constructive heuristics based on principal-components approximation for solving the GP problem with fixed but not necessarily equal part sizes. Low-rank approximation of the connection matrix is motivated by an observation made in this paper that the GP problem on low-rank matrices can be solved in polynomial time. Justification for such an approximation is based on another remarkable observation, that there exists a *rank-one* positive semidefinite matrix Q such that an optimal partition with respect to Q is also optimal with respect to the original connection matrix. The approximation used is based on principal components and requires computing a few largest eigenvalues and corresponding eigenvectors of an $n \times n$ matrix for a graph on n vertices. New lower bounds on the cost of an optimal partition are also obtained from these eigenvalues and eigenvectors, and these bounds are shown to be significantly better than the previously known Donath–Hoffman lower bounds. The quality of the performance of these heuristics and the corresponding lower bounds improves quite significantly on average as the size of the graph increases or the density factor approaches 1.0. One of the heuristics of this paper, namely, GP-2D.0, is able to produce partitions with cost within 10% of a lower bound LB_2 for graphs on 300 or more vertices and with density $\gamma \geqq 0.1$. A much faster heuristic is GP-1D.0, which produces partitions with cost within 5% of those produced by GP-2D.0. In case further improvement in cost is desired, we suggest using the partition generated by GP-1D.0 as an initial partition for iterative improvement techniques, such as the Kernighan–Lin method [24], or even stochastic search methods, such as simulated annealing [25], [26], or stochastic evolution [27].

Appendix.

Proof of Proposition 8 (*for the graph bisectioning problem*, $\text{LB}_1 \geqq \text{LB}_{\text{DH}}$). In the special case of bisectioning, LB_{DH} reduces to $\text{LB}_{\text{DH}} = \frac{1}{2}\underline{\mathbf{1}}^T A \underline{\mathbf{1}} - (n/4)(\mu_1 + \mu_2)$. With $n_1 = n_2 = n/2$ in (8.18), we have $\sum_{i=1}^{l} \hat{x}_i = n/4$. Hence

$$\text{LB}_1 = \frac{1}{4}\underline{\mathbf{1}}^t A \underline{\mathbf{1}} - \sum_{i=1}^{l} \lambda_i \hat{x}_i \geqq \frac{1}{4}\underline{\mathbf{1}}^t A \underline{\mathbf{1}} - \lambda_1 \left[\sum_{i=1}^{l} \hat{x}_i \right] = \frac{1}{4}\underline{\mathbf{1}}^t A \underline{\mathbf{1}} - \frac{n}{4}\lambda_1 .$$

We will now prove that $\mu_1 + \mu_2 \geqq (1/n)\underline{\mathbf{1}}^t A \underline{\mathbf{1}} + \lambda_1$, which will establish the result. To this end, let $\{\underline{\mathbf{w}}_i\colon A\underline{\mathbf{w}}_i = \mu_i \underline{\mathbf{w}}_i,\ i = 1, 2, \ldots, n\}$ denote an *orthonormal* set of eigenvectors of A. Similarly, let $\{\underline{\mathbf{e}}_i\colon M\underline{\mathbf{e}}_i = \lambda_i\underline{\mathbf{e}}_i, i = 1, 2, \ldots, n\}$ denote an *orthonormal* set of eigenvectors of M, where $\underline{\mathbf{e}}_n = (1/\sqrt{n})\underline{\mathbf{1}}$. Since the two sets of eigenvectors of A and M are individually orthonormal basis sets in $\mathbb{R}^n$, we can express the eigenvectors of M as a linear combination of the eigenvectors of A as $\underline{\mathbf{e}}_i = \sum_{j=1}^n \xi_{ij}\underline{\mathbf{w}}_j$ for each $i = 1, 2, \ldots, n$, where the scalars $\xi_{ij} = \underline{\mathbf{e}}_i^t\underline{\mathbf{w}}_j$ can be easily shown to satisfy

$$\text{(A.1)} \qquad \sum_{j=1}^{n} \xi_{ij}^2 = 1 \quad \text{for all} \quad 1 \leqq i \leqq n$$

and

$$\text{(A.2)} \qquad \sum_{i=1}^{n} \xi_{ij}^2 = 1 \quad \text{for all} \quad 1 \leqq j \leqq n.$$

Moreover, for each $i = 1, 2, \ldots, n$ we have $\underline{\mathbf{e}}_i^t A\underline{\mathbf{e}}_i = \sum_{j=1}^n \xi_{ij}^2\mu_j$. Now, $\underline{\mathbf{e}}_1^t A\underline{\mathbf{e}}_1 = \underline{\mathbf{e}}_1^t PAP\underline{\mathbf{e}}_1 = \lambda_1$ since $\underline{\mathbf{e}}_1^t\underline{\mathbf{1}} = 0$ implies $P\underline{\mathbf{e}}_1 = \underline{\mathbf{e}}_1$. Furthermore, $\underline{\mathbf{e}}_n^t A\underline{\mathbf{e}}_n = (1/n)\underline{\mathbf{1}}^t A\underline{\mathbf{1}}$. Combining these yields

$$\begin{aligned} \frac{1}{n}\underline{\mathbf{1}}^t A\underline{\mathbf{1}} + \lambda_1 &= \sum_{j=1}^{n} (\xi_{nj}^2 + \xi_{1j}^2)\mu_j \\ &\leqq (\xi_{n1}^2 + \xi_{11}^2)\mu_1 + \left[\sum_{j=2}^{n} (\xi_{nj}^2 + \xi_{1j}^2)\right]\mu_2 \\ &= (\xi_{n1}^2 + \xi_{11}^2)\mu_1 + (2 - \xi_{n1}^2 + \xi_{11}^2)\mu_2 \\ &\leqq \mu_1 + \mu_2, \end{aligned}$$

where the first inequality follows from $\mu_1 \geqq \mu_2 \geqq \cdots \mu_n$, the second equality follows from (A.1), and the last inequality holds because $0 \leqq \xi_{n1}^2 + \xi_{11}^2 \leqq 1$ from (A.2) implies that $(\xi_{n1}^2 + \xi_{11}^2)\mu_1 + (1 - \xi_{n1}^2 + \xi_{11}^2)\mu_2$ is a convex combination of μ_1 and μ_2. This completes the proof of this proposition. □

Proof of Proposition 9 (*if A has equal row sums, then* $\mathrm{LB}_1 \geqq \mathrm{LB}_{\mathrm{DH}}$). Since $A\underline{\mathbf{1}} = \mu_1\underline{\mathbf{1}}$, we have $r^* = \mu_1 n_1/n$ and $\eta = \mu_1/n$. Using (8.13) of Proposition 7 (with $p = 1$) and (8.15), we get

$$\mathrm{LB}_1 = \frac{n_1 n_2}{n}\mu_1 - \sum_{i=1}^{l} \lambda_i\hat{x}_i = \frac{n_1 n_2}{n}\mu_1 - \sum_{i=1}^{l} \mu_{i+1}\hat{x}_i.$$

An examination of (8.12) shows that $\sum_{i=1}^{l} \hat{x}_i = n_1 n_2/n$. Thus $\mathrm{LB}_1 \geqq (n_1 n_2/n)(\mu_1 - \mu_2)$. On the other hand, using $A\underline{\mathbf{1}} = \mu_1\underline{\mathbf{1}}$ in (8.14) implies that $\mathrm{LB}_{\mathrm{DH}} = (n_2/2)(\mu_1 - \mu_2)$. Since $n_1 \geqq n_2$ and $n_1 + n_2 = n$, the desired inequality follows. □

REFERENCES

[1] M. R. Garey, D. S. Johnson, and L. Stockmeyer, *Some simplified* NP-*complete graph problems*, Theoret. Comput. Sci., 1 (1976), pp. 237-267.

[2] T. C. Hu and E. S. Kuh, VLSI *Circuit Layout: Theory and Design*, IEEE Press, New York, 1985.

[3] S. N. Bhatt and F. T. Leighton, *A framework for solving* VLSI *graph problems*, J. Comput. System Sci., 28 (1984), pp. 300-343.

[4] C. R. Rao, *The use and interpretation of principal component analysis in applied research*, Sankhyā Ser. A, 26 (1964), pp. 329-358.

[5] K. S. Arun and V. B. Rao, *A new principal components algorithm for graph partitioning with applications in* VLSI *placement problems*, in Proc. 20th Annual Asilomar Conference on Signals, Systems, and Computers, IEEE Computer Society, Washington, DC, 1986, pp. 725–729.

[6] J. D. Ullman, *Computational Aspects of* VLSI. Computer Science Press, Rockville, MD, 1984.

[7] J. Frankle and R. M. Karp, *A graph partitioning algorithm based on spectral analysis*, in Proc. SRC Topical Research Conference, Berkeley, CA, 1985.

[8] V. B. Rao and K. S. Arun, *Two-way graph partitioning by principal components*, in Proc. 1990 IEEE International Symposium on Circuits and Systems, Institute of Electrical and Electronic Engineers, New York, 1990, pp. 2877–2880.

[9] K. S. Arun and V. B. Rao, *Rank reduction in graph partitioning*, in Proc. 1991 IEEE International Conference on Acoustics, Speech, and Signal Processing, Institute of Electrical and Electronic Engineers, New York, 1991, pp. 3297–3300.

[10] K. S. Arun and V. B. Rao, *New heuristics and lower bounds for graph partitioning*, in Proc. 1991 IEEE International Symposium on Circuits and Systems, Institute of Electrical and Electronic Engineers, New York, 1991, pp. 1172–1175.

[11] J. C. Gower, *Properties of Euclidean and non-Euclidean distance matrices*, Linear Algebra Appl., 67 (1985), pp. 81–97.

[12] W. Glunt, T. L. Hayden, S. Hong, and J. Wells, *An alternating projections algorithm for computing the nearest Euclidean distance matrix*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 589–600.

[13] A. Pothen, H. D. Simon, and K. P. Liou, *Partitioning sparse matrices with eigenvectors of graphs*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 430–452.

[14] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *Data Structures and Algorithms*, Addison-Welsey, Reading, MA, 1983.

[15] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[16] ———, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[17] K. Pearson, *On lines and planes of closest fit to systems of points in space*, Philos. Mag. (6th Ser), 2 (1901), pp. 559–572.

[18] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1976.

[19] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[20] B. Fruchter, *Introduction to Factor Analysis*, Van Nostrand, New York, 1954.

[21] L. L. Thurstone, *Multiple Factor Analysis*, University of Chicago Press, Chicago, 1947.

[22] W. E. Donath and A. J. Hoffman, *Lower bounds for the partitioning of graphs*, IBM J. Res. Develop. 17 (1973), pp. 420–425.

[23] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[24] B. W. Kernighan and S. Lin, *An efficient heuristic procedure for partitioning graphs*, Bell Syst. Tech. J., 49 (1970), pp. 291–307.

[25] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science, 20 (1983), pp. 671–680.

[26] J. Lam and J. M. Delosme, *Simulated annealing*: *A fast heuristic for some generic layout problems*, in Proc. IEEE International Conference on Computer-Aided Design, Institute of Electrical and Electronics Engineers, New York, 1988, pp. 510–513.

[27] Y. G. Saab and V. B. Rao, *Combinatorial optimization by stochastic evolution*, IEEE Trans. Computer-Aided Design, 10 (1991), pp. 525–535.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1016–1042, October 1993

007

FAST TRIANGULAR FACTORIZATION AND INVERSION OF HERMITIAN, TOEPLITZ, AND RELATED MATRICES WITH ARBITRARY RANK PROFILE*

DEBAJYOTI PAL† AND THOMAS KAILATH‡

Abstract. A fast procedure for computing a "modified" triangular factorization and inverse of Hermitian Toeplitz and quasi-Toeplitz (matrices congruent in a certain sense to Toeplitz matrices) matrices is presented. A modified triangular factorization is an LDL^* factorization where L is lower triangular with unit diagonal entries and D is a block diagonal matrix with possibly varying block sizes; only matrices with all leading minors nonzero, often called strongly regular, will always have a purely diagonal and nonsingular D matrix. For the matrices studied herein, the diagonal blocks also have a particular quasi-Toeplitz structure.

The algorithms are obtained by extending a generating function approach of Lev-Ari and Kailath [*Operator Theory*: *Adv. Appl.*, 18 (1986), pp. 301–324.] for matrices with a generalized displacement structure. A particular application of the result is a fast method of computing the rank profile and inertia of the matrices involved.

Key words. triangular factorization, Schur complement, Toeplitz, quasi-Toeplitz, inversion, zero minors, rank profile, inertia

AMS subject classifications. 15A06, 15A09, 15A23, 15A57, 65F05

1. Introduction. Fast $O(n^2)$ algorithms for linear algebra problems associated with Toeplitz matrices have been studied for many years now, starting perhaps with the classical work of Levinson [Lev47] on the solution of linear equations with a Toeplitz coefficient matrix; Levinson's algorithm was $O(n^2)$ for an $n \times n$ Toeplitz matrix as compared to $O(n^3)$ for an arbitrary $n \times n$ matrix. Elaborations and extensions were supplied by many authors, especially Durbin [Dur60], Trench [Tre64], Zohar [Zoh69], and others. Durbin [Dur60] studied the particular case of Toeplitz matrix systems with the right-hand side being proportional to the first unit vector, $[1\ 0\ 0\ \cdots\ 0]^*$. The relation of Durbin's recursions to those for the Szegö orthogonal polynomials was perhaps first made by Whittle [Whi63, p. 36], and by Trench [Tre64]. Trench in fact essentially noted that the Christoffel–Darboux formula for the Szegö polynomials (orthogonal on the unit circle) gave a fast recursion for finding the elements of the *inverse* of a symmetric positive definite Toeplitz matrix. The "summed" form of these recursions gives a result that is a special case of a celebrated formula of Gohberg and Semencul [GS72]; for more discussion and further references see the survey papers of Kailath, Vieira, and Morf [KVM78] and Kailath [Kai87, Appendix C]. It was also known to many authors (see, e.g., [Bur75]) that arranging the successive solutions of the normal equations in a lower triangular matrix essentially gave (apart from scaling) the lower triangular factor in a so-called UDL (upper diagonal lower) triangular factorization of the *inverse* of the coefficient matrix.

In many problems, it is of interest to determine the triangular factors of the matrix itself rather than its inverse. Fast algorithms for doing this can be inferred after some algebra from the Levinson Algorithm as, e.g., in the survey Kailath [Kai87, Appendix

* Received by the editors November 13, 1989; accepted for publication (in revised form) March 11, 1992. This work was supported in part by U.S. Army Research Office contract DAAL03-86-K-0045 and the Strategic Defense Initiative Organization/Innovative Science & Technology, managed by Army Research Office contract DAAL03-87-K-0033.

† Information Systems Research Laboratory, AT&T Bell Laboratories, 101 Crawfords Corner Road, Holmdel, New Jersey 07733 (debu@vax135.att.com).

‡ Information Systems Laboratory, Stanford University, Stanford, California 94305 (tk.isl.stanford.edu).

B]. The first explicit solutions of the direct (so-called Cholesky) factorization problem for Toeplitz matrices were given in different contexts and in somewhat different ways by Bareiss [Bar69]; Morf [Mor70], [Mor74]; Rissanen [Ris73]; and Le Roux and Gueguen [RG77]. It was noticed that all these algorithms were essentially related to a recursive algorithm of Schur [Sch17] for checking if a power series is analytic and bounded in the unit disc. Schur's work was perhaps first introduced into this circle of problems by Dewilde, Vieira, and Kailath [DVK78]; further study of Schur's paper by Lev-Ari and Kailath [LK84] showed that the same algorithm also provides the triangular factorization of any positive definite matrix of the form

$$(1.1) \qquad Q = L(\alpha)L^*(\alpha) - L(\beta)L^*(\beta),$$

where $L(\alpha)$ is a lower triangular Toeplitz matrix with first column α and $(^*)$ denotes Hermitian transpose. Such matrices will be called quasi-Toeplitz (QT), in part because of the identity (first given by Schur [Sch17]),

$$(1.2) \qquad Q = L(\alpha)L^*(\alpha) - L(\beta)L^*(\beta) = L(\alpha+\beta)TL^*(\alpha+\beta),$$

where T is a Toeplitz matrix (said to be *congruent* to Q). It may be of interest to point out a physical interpretation via lossless transmission lines of the Schur factorization algorithm; see Kailath, Bruckstein, and Morgan [KBM86].

After this lengthy review it must be noted that all of the above results are for positive definite Toeplitz matrices, or at least for Toeplitz matrices with all nonzero leading minors—so-called *strongly nonsingular* or *strongly regular* matrices. Successful attempts to generalize the Levinson algorithm, and consequently to compute the inverse matrix, to nonstrongly regular Toeplitz matrices were made by Heinig and Rost [HR84] and later by Delsarte, Genin, and Kamp [DGK85] and in a different way (using generalized three-term recursions) by Pombra, Lev-Ari, and Kailath [PLK88]. However, there appears to be no prior results on the direct triangular factorization of nonstrongly regular Toeplitz matrices despite scattered attempts by various researchers. We should mention here that a strict triangular factorization is not possible unless all the leading minors are nonzero. What can be achieved otherwise is a modified triangular factorization of the form

$$(1.3) \qquad M = LD_BL^*,$$

where L is lower triangular with unit diagonal, while D_B is a block diagonal matrix, with possibly varying block sizes. In this paper we will present a recursive $O(n^2)$ algorithm for such a factorization. Furthermore, following the ideas of Chun [Chu89], we will develop fast $O(n^2)$ algorithms for factorizing certain block Toeplitz matrices that will allow us to obtain alternative $O(n^2)$ algorithms to those mentioned above for solving linear equations and obtaining the inverse of the coefficient matrix.

A characteristic of all these new algorithms is that the coefficients in the recursions are all computed without invoking inner products, a fact that sets them apart from all Levinson-related algorithms. The absence of inner products makes parallel computation more feasible in the sense that with $O(n)$ processors, the computation time for the direct factorization methods can be reduced to $O(n)$, assuming unit time for each elementary computation; because of the inner products, the corresponding number for the Levinson-type algorithms is $O(n \log n)$.

To close this introduction we should mention a number of papers that solve the problems for Toeplitz matrices by converting them to Hankel matrices (by reversing the order of the columns); see, e.g., Brent, Gustavson, and Yun [BGY80]; Sugiyama [Sug86]; and Labahn, Choi, and Cabay [LCC90]. The point is that the corresponding problem for Hankel matrices appears to be easier to solve, with solutions going back to Berlekamp

[Ber68], Massey [Mas69], Gragg [Gra74], and many others. However, recursive solutions (in the size of the Toeplitz matrix) are apparently ruled out by such an approach. Our paper focuses only on recursive algorithms, which are important in various problems. For completeness we should also mention a paper of Gover and Barnett [GB85], who use certain so-called r-Toeplitz matrices; however, the algorithm is not efficient in terms of the amount of computation and requires a lot of $O(r^3)$ steps and a considerable effort to find the right choices of r.

In the next section we explain the general (Jacobi) triangular factorization procedure for an arbitrary matrix. In § 3 we show how the structure of Toeplitz and QT matrices can be exploited to obtain a fast recursive procedure for computing a modified triangular factorization of Toeplitz and QT matrices. It is shown that the diagonal blocks are also QT. Following Lev-Ari and Kailath [LK86] we introduce a "generating function" representation of Hermitian matrices and derive all our results in the form of certain polynomial recursions. In § 4, we consider the inertia properties and show that the signature of the block diagonal entries is always zero. In § 5 we introduce the notions of *block generating functions* and *block quasi-Toeplitz* (BQT) matrices. It is shown that inverse of a QT matrix can be obtained by computing a block Schur complement of a certain associated BQT matrix. Next we discuss the so-called *admissibility conditions* and derive polynomial recursions for Schur complementation of BQT matrices with regular and arbitrary rank profiles. At this point we discuss the transmission line interpretation of these polynomial recursions and indicate that this procedure works under all circumstances, including those where a Gohberg–Semencul formula does not hold good. We end this section with an example. In § 6 we show how to adopt the procedure of § 5 toward obtaining solutions to linear equations with a nonsingular QT coefficient matrix and an arbitrary right-hand side. Section 7 contains some concluding remarks.

2. Triangular factorization of Hermitian matrices. Strongly regular Hermitian matrices $M \in C^{n \times n}$, can be factored as

$$M = LDL^*, \tag{2.1}$$

where L is a lower triangular matrix with unit diagonal elements, D is a nonsingular diagonal matrix, and the superscript * denotes Hermitian transpose.

Let us denote the columns of matrix L by $\{\mathbf{l}_i,\ i = 1, 2, \ldots, n\}$ and the diagonal elements of the matrix D by $\{d_i\}$. Then we can write

$$M = LDL^* = \sum_{i=1}^{n} d_i \mathbf{l}_i \mathbf{l}_i^*, \tag{2.2}$$

which suggests the following recursive computational procedure:

$$M_{i+1} = M_i - d_i \mathbf{l}_i \mathbf{l}_i^*; \quad M_1 = M, \tag{2.3}$$

$$d_i = \mathbf{e}_i^* M_i \mathbf{e}_i \quad \text{and} \quad \mathbf{l}_i = M_i \mathbf{e}_i d_i^{-1}, \tag{2.4}$$

where $\mathbf{e}_i$ is the unit vector with nonzero entry at the ith position.

The above procedure is the celebrated Jacobi procedure (see [LK86]). Equations (2.2)–(2.4) show that the first $i - 1$ rows and columns of M_i contain only zero entries, d_i is the (i, i)th element of M_i, and $d_i \mathbf{l}_i$ is the ith column of M_i.

So M_i has the form

$$M_i = \left[\begin{array}{c|c} O & O \\ \hline O & P_i \end{array}\right],$$

where P_i is called the Schur complement of the matrix M with respect to the leading

$(i-1) \times (i-1)$ principal submatrix; note that P_{i+1} is the Schur complement of the (2.1) entry of P_i, i.e., we compute one Schur complement at every step of the triangular factorization procedure (2.3)–(2.4).

Singularities. Given a real symmetric matrix M it will be impossible to compute the Schur complement of the (2.1) element if and only if the (2.1) element is zero. The following are the possibilities: (i) $M = 0$; (ii) $M \neq 0$, but the first row and column of M are identically zero; and (iii) $M \neq 0$, the first row and column of M are also not identically zero, but $m_{11} = 0$.

In case (i) finding a triangular factorization is trivial. In case (ii) there is no need to find the Schur complement of m_{11}. We simply choose the first column of L to be $\mathbf{e}_1$, the first unit vector and $d_1 = 0$; then resume the triangular factorization procedure. In case (iii) we must find the block Schur complement of the smallest nonsingular block available at the top left corner of the matrix M. This will lead to a *modified* triangular factorization $M = LDL^*$, where $L \in C^{n \times n}$ is a lower triangular matrix with unit diagonal elements and $D \in C^{n \times n}$ is a Hermitian matrix with scalar as well as block entries along the diagonal, while all the other entries are zero. The matrix D is nonsingular if and only if M is. Such a factorization always exists; however, it is nonunique (see also Bunch, Kauffman, and Parlett [BKP76]).

3. Modified triangular factorization of Toeplitz and quasi-Toeplitz matrices.

3.1. Generating functions. Since triangular factorization is nested, factorization of the leading principal submatrices does not depend on the actual size of the matrix. Hence we can consider, without loss of generality, that the matrix under consideration is in fact semi-infinite. Following [LK86], we will associate such a matrix M with a "generating" function

$$M(z,w) = [1 \quad z \quad z^2 \quad \cdots \quad \cdots]M[1 \quad w \quad w^2 \quad \cdots \quad \cdots]^*. \tag{3.1}$$

A Toeplitz matrix of order n $(=1, 2, \ldots)$ has the form $T_{n-1} = [t_{i-j}]_{i,j=0}^{n-1}$, where the $\{t_k\}$ are arbitrary $(k = -(n-1), \ldots, -1, 0, 1, \ldots, (n-1))$. If T_{n-1} is Hermitian, then the entries $\{t_k\}_{-(n-1)}^{(n-1)}$ must satisfy $t_{-k} = t_k^*$ for $0 \leqq k \leqq n-1$. The semi-infinite extension $T_{n-1,\infty}$ of a finite Hermitian Toeplitz matrix T_{n-1} is the banded semi-infinite Toeplitz matrix with the first row $[t_0^* \quad t_1^* \quad \cdots \quad \cdots \quad t_{(n-1)}^* \quad 0 \quad 0 \quad \cdots \quad \cdots]$, and its generating function can be seen to be

$$T_{n-1,\infty}(z,w) = \frac{t(z) + t^*(w)}{1 - zw^*}, \tag{3.2}$$

where

$$t(z) = \frac{1}{2}t_0 + t_1 z + \cdots + t_{n-1}z^{n-1}. \tag{3.3}$$

Similarly, the generating function $Q(z, w)$ of the semi-infinite extension of the QT matrix Q in (1.1) can be represented as

$$Q(z,w) = \frac{\alpha(z)\alpha^*(w) - \beta(z)\beta^*(w)}{1 - zw^*}, \tag{3.4}$$

where $\alpha(z)$ and $\beta(z)$ are polynomials (power series) in z, say

$$\alpha(z) = \sum_{k=0}^{m} \alpha_k z^k \quad \text{and} \quad \beta(z) = \sum_{k=0}^{q} \beta_k z^k. \tag{3.5}$$

THEOREM 3.1. *The Schur complement of the* 1×1 *leading principal submatrix of* $T_{n-1,\infty}$ *is QT provided* $t_0 \neq 0$.

Proof. If $t_0 = 0$, then the Schur complement does not exist. So we assume $t_0 \neq 0$. The generating function $T^c_{n-1,\infty}(z, w)$ of this Schur complement is

(3.6)

$$(zw^*)T^c_{n-1,\infty}(z, w) = T_{n-1,\infty}(z, w) - T_{n-1,\infty}(z, 0)[T_{n-1,\infty}(0, 0)]^{-1}T_{n-1,\infty}(0, w).$$

Rearranging (3.3),

$$T_{n-1,\infty}(z, w) = \frac{(t(z) + \frac{1}{2}t_0)(t(w) + \frac{1}{2}t_0)^* - (t(z) - \frac{1}{2}t_0)(t(w) - \frac{1}{2}t_0)^*}{(1 - zw^*)t_0},$$

which implies that

$$T^c_{n-1,\infty}(z, w) = \frac{(t(z) + \frac{1}{2}t_0)(t(w) + \frac{1}{2}t_0)^* - t_1(z)t_1^*(w)}{(1 - zw^*)t_0},$$

where we have defined

$$t_1(z) = \sum_{k=0}^{n-2} t_{k+1}z^k. \qquad \square \tag{3.7}$$

Since at least one Schur complement is evaluated at every step of triangular factorization, a question naturally arises about the structure of the Schur complement of a QT matrix.

THEOREM 3.2. *The Schur complement of the* 1×1 *leading principal submatrix of a QT matrix is also QT provided* $Q(0, 0) \neq 0$.

Proof. The generating function $Q^c(z, w)$ of the Schur complement is given by

$$(zw^*)Q^c(z, w) = Q(z, w) - Q(z, 0)[Q(0, 0)]^{-1}Q(0, w), \tag{3.8}$$

where $Q(z, w)$ is as in (3.4), and we have assumed that $Q(0, 0) \neq 0$. In this case some algebra will show that

$$Q^c(z, w) = \frac{[\alpha(z)\alpha^*(0) - \beta(z)\beta^*(0)][\alpha(0)\alpha^*(w) - \beta(0)\beta^*(w)] - \beta_1(z)\beta_1^*(w)}{(|\alpha(0)|^2 - |\beta(0)|^2)(1 - zw^*)},$$

where

$$\beta_1(z) = [\alpha(z)\beta(0) - \beta(z)\alpha(0)]/z. \tag{3.9}$$

Clearly, $\beta_1(z)$ is a polynomial since the numerator in (3.9) has a root at the origin. This completes the proof of the theorem, since $Q^c(z, w)$ is clearly QT. $\square$

3.2. What can we do when $Q(0, 0) = 0$. The following are the possibilities (see § 2).

Case (i). $Q(z, w) = 0$, which will happen if and only if $\alpha(z) = e^{j\phi}\beta(z)$, for some $\phi \in [0, 2\pi]$.

Case (ii). $Q \neq 0$, but the first row and the column are identically zero, i.e., $Q(z, w) \neq 0$, but $Q(z, 0) = 0 = Q(0, w)$. This is equivalent to having $\alpha(0) = 0 = \beta(0)$.

Case (iii). Q is not zero, neither the first column nor the first row is zero, but the 1×1 entry of Q is zero, i.e., $Q(0, 0) = 0$, $Q(z, 0) \neq 0$, $Q(0, w) \neq 0$, and $Q(z, w) \neq 0$. This implies that

$$|\alpha(0)|^2 = |\beta(0)|^2; \quad \alpha(0) \neq 0, \quad \beta(0) \neq 0, \quad \text{and} \quad \alpha(z) \neq e^{j\phi}\beta(z) \quad \text{for } any\, \phi \in [0, 2\pi].$$

Carrying on with the triangularization is simple in Cases (i) and (ii). However, Case (iii) is more difficult to deal with. To see what is happening here, it will help to first rewrite the expression (3.4) for $Q(z, w)$ using the so-called immittance variables

$$(3.10) \qquad \begin{cases} \sqrt{2}a(z) = \alpha(z) - (\beta^*(0)/\alpha^*(0))\beta(z), \\ \sqrt{2}b(z) = \alpha(z) + (\beta^*(0)/\alpha^*(0))\beta(z), \end{cases}$$

to get

$$(3.11) \qquad Q(z, w) = \frac{a(z)b^*(w) + b(z)a^*(w)}{1 - zw^*}.$$

The condition $|\alpha(0)|^2 = |\beta(0)|^2$ can now be restated as follows.

THEOREM 3.3. *The polynomial $a(z)$ must have at least one root at the origin and if $a(z)$ has t roots at the origin, the first nonsingular leading principal submatrix of Q must be of dimension $2t$.*

Proof. Since

$$a(z) = \alpha(z) - \frac{\beta^*(0)}{\alpha^*(0)}\beta(z); \quad \beta(0) \neq 0,$$

the first claim is verified immediately. Now suppose $a(z)$ has t roots at the origin. Then $a_j = 0$, $0 \leqq j \leqq t - 1$, which implies that

$$(3.12) \qquad a(z) = z^t a_t(z), \quad a_t(z) = \sum_{j=0}^{p-t} a_{j+t} z^j.$$

Therefore,

$$(3.13) \qquad Q(z, w) = \frac{z^t a_t(z)b^*(w) + (w^*)^t a_t^*(w)b(z)}{1 - zw^*}.$$

Now introduce the Taylor series for the rational function $g(z)$ defined as

$$(3.14) \qquad \frac{b(z)}{a_t(z)} = g(z) = \sum_{k=0}^{\infty} g_k z^k.$$

Then we can write

$$Q(z, w) = a_t(z)\tilde{G}(z, w)a_t^*(z),$$

where

$$(3.15) \qquad \tilde{G}(z, w) = \frac{[z^t g^*(w) + (w^*)^t g(z)]}{1 - zw^*}.$$

Clearly, $z^t g^*(w)$ represents a semi-infinite matrix whose $(t + 1)$th row from the top is $[g_0^*, g_1^*, g_2^* \cdots \cdots]$, and all the other rows are zeros; $(w^*)^t g(z)$ represents the Hermitian transpose of this semi-infinite matrix. Thus the arrangement of the zero and nonzero elements of the matrix $\tilde{G}$, associated with $\tilde{G}(z, w)$, must have the form

$$(3.16) \qquad \tilde{G} = \left[\begin{array}{c|cccccc} O & \multicolumn{1}{c|}{U^*} & O & \cdots & \cdots & & \\ \hline U & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ O & \cdot & \cdot & \hat{G} & \cdot & \cdot & \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right],$$

where U is a $t \times t$ upper triangular Toeplitz matrix, and $\hat{G}$ is a semi-infinite Hermitian Toeplitz matrix. Clearly, the entry on the diagonal of U is g_0^*. Since $b(0) \neq 0$ implies $g_0 \neq 0$, U is full rank. Therefore, its $2t \times 2t$ first leading principal submatrix must be full rank. All other $j \times j$, $1 \leqq j \leqq 2t-1$, leading principal submatrices of $\tilde{G}$ can be verified to be rank deficient by inspection. This completes the proof of the theorem. □

LEMMA 3.1. Formula for $\tilde{G}$. *The first row of the $t \times t$ upper triangular Toeplitz matrix U is $[g_0^* \cdot g_1^*, \dots, g_{t-1}^*]$, while the first column of the Hermitian Toeplitz matrix $\hat{G}$ is $[(\hat{g}_0 + \hat{g}_0^*), \hat{g}_1^*, \dots, \dots]$, where*

$$\hat{g}_i = \begin{cases} g_t & i = 0, \\ g_{t+i} + g_{t-i}^* & 1 \leqq i \leqq t, \\ g_{t+i} & i \geqq t+1. \end{cases}$$

Proof. The proof is obtained by direct coefficient matching. □

THEOREM 3.4. *The block Schur complement of the first $2t \times 2t$ leading principal submatrix of $\tilde{G}$ is $\hat{G}$.*

Proof. Note that

$$\hat{G} = \left[\begin{array}{c|ccccc} \hat{G}_0 & \hat{G}_1^* & \cdots & \cdots & \cdots \\ \text{—} & \cdot & \cdot & \cdot & \cdot \\ \hat{G}_1 & \cdot & \cdot & \cdot & \cdot \\ \vdots & \cdot & \cdot & \hat{G} & \cdot \\ \vdots & \cdot & \cdot & \cdot & \cdot \\ \vdots & \cdot & \cdot & \cdot & \cdot \end{array}\right], \tag{3.17}$$

where the $t \times t$ Toeplitz matrices $\{\hat{G}_i\}_0^\infty$ are formed from the coefficients of the first t rows and columns of $\hat{G}$ by proper partitioning. Hence the Schur complement $\tilde{G}_c$ under consideration is

$$\tilde{G}_c = \hat{G} - \begin{bmatrix} 0 & \hat{G}_1 \\ 0 & \hat{G}_2 \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} 0 & U^* \\ U & \hat{G}_0 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 & \cdots & \cdots \\ \hat{G}_1^* & \hat{G}_2^* & \cdots & \cdots \end{bmatrix} = \hat{G}. \qquad \square$$

However, $\hat{G}$ can be obtained directly from $\tilde{G}$ by removing U and U^* from it, i.e.,

$$\left[\begin{array}{c|cccccc} O & O & \vert & O & \cdots & \cdots \\ \hline O & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot \\ O & \cdot & \cdot & \hat{G} & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right] \tag{3.18}$$

$$= \tilde{G} - \left[\begin{array}{c|cccccc} O & U^* & \vert & O & \cdots & \cdots \\ \hline U & O & O & \cdot & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot \\ O & O & \cdot & O & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdots & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right].$$

Thus the generating function of $\tilde{G}_c$ can be represented as follows:

(3.19)

$$(zw^*)^t G^c(z, w) = \tilde{G}(z, w) - \frac{[\sum_{j=0}^{t-1} [g_j^* z^{t-j} + g_j(w^*)^{t-j}](zw^*)^j[1 - (zw^*)^{t-j}]]}{1 - zw^*} .$$

Clearly, the above process of "removing U^* and U" from $\tilde{G}$ can be carried out in t steps such that at every step one of the diagonals (super diagonals) of U and a corresponding diagonal (subdiagonal) of U^* are removed.

Since $Q_{2t}^c(z, w) = a(z)\tilde{G}_c a^*(w)$, it should be possible to obtain Q_{2t}^c in t scalar steps, also.

Observe the first step:

$$(3.20)\qquad \begin{aligned}&(zw^*)^{t-1}\tilde{G}^c(z, w)\\ &= -\left[\sum_{j=0}^{t-2} [g_{j+1}^* z^{t-j-1} + g_{j+1}(w^*)^{t-j-1}](zw^*)^j[1 - (zw^*)^{t-j-1}]/(1 - zw^*)\right]\\ &\quad + Q_1^c(z, w)/a(z)a^*(w),\end{aligned}$$

where

$$(3.21)\qquad \begin{aligned}&zw^*Q_1^c(z, w)\\ &\quad = Q(z, w) - a_t(z)[(g_0^* z^t + g_0(w^*)^t)(1 - (zw^*)^t)/(1 - zw^*)]a_t^*(w).\end{aligned}$$

Since

$$Q(z, w) = \frac{z^t a_t(z)b^*(w) + (w^*)^t a_t^*(w)b(z)}{1 - zw^*},$$

$$\begin{aligned}&Q(z, w) - a_t(z)[(g_0^* z^t + g_0(w^*)^t)(1 - (zw^*)^t)/(1 - zw^*)]a_t^*(w)\\ &\qquad = \frac{\begin{aligned}&z^t a_t(z)[b^*(w) - g_0^* a_t^*(w) - (w^*)^{2t} g_0 a_t^*(w)]\\ &\quad + (w^*)^t a_t^*(w)[b(z) - g_0 a_t(z) - z^{2t} g_0^* a_t]\end{aligned}}{1 - zw^*} .\end{aligned}$$

Since $g_0 = b(0)/a_t(0)$, $\lim_{z\to 0} b(z) - g_0 a_t(z) - z^{2t} g_0^* a_t(z) = 0$. Thus,

$$(3.22)\qquad Q_1^g(z, w) = \frac{z^{t-1} a_t(z)b_1^*(w) + (w^*)^{t-1} a_t^*(w)b_1(z)}{1 - zw^*},$$

where

$$zb_1(z) = b(z) - g_0 a_t(z) + z^{2t} g_0^* a_t(z) \quad \text{and} \quad g_0 = \lim_{z\to 0} b(z)/a_t(z).$$

This suggests the following result.

LEMMA 3.3. The recursive form for $Q_{2t}^c(z, w)$ is as follows.

$$(3.23)\qquad Q_{2t}^c(z, w) = \frac{a_t(z)b_t^*(w) + b_t(z)a_t^*(w)}{1 - zw^*},$$

where $b_t(z)$ is defined by the recursion

$$(3.24)\qquad \begin{aligned}&b_0(z) = b(z),\\ &zb_{j+1}(z) = b_j(z) - \tau_j a_t(z) + \tau_j^* z^{2(t-j)} a_t(z); \quad \tau_j = b_j(0)/a_t(0).\end{aligned}$$

Proof. Define the sequence

$$Q_j^g(z, w) = \frac{z^{t-j}a_t(z)b_j^*(w) + (w^*)^{t-j}a_t^*(w)b_j(z)}{1 - zw^*}, \quad j = 1, \dots, t,$$

where $\{b_j(z)\}_{j-1}^t$ is defined by the recursion (3.24). Then

$$Q_j^\tau(z, w) - a_t(z)a_t^*(w)[\tau_j^* z^{t-j} + \tau_j(w^*)^{t-j}][1 - (zw^*)^{t-j}]/(1 - zw^*)$$
$$= (zw^*)Q_{j+1}^\tau(z, w).$$

It is clear that there is a relationship between the $\{\tau_i\}$ and the coefficients $\{g_i\}$ of the power series expansion of $g(z)$. We will repeat (3.14) for convenience:

$$\frac{b(z)}{a_t(z)} = g(z) = \sum_{k=0}^{\infty} g_k z^k,$$

whose matrix version is

$$\text{(3.25)} \quad \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_r \\ 0 \\ 0 \\ \vdots \end{bmatrix} = \begin{bmatrix} a_t & 0 & 0 & \cdots & \cdots & g_0 \\ a_{t+1} & a_t & 0 & \ddots & \ddots & g_1 \\ a_{t+2} & a_{t+1} & a_t & 0 & \ddots & g_2 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ a_r & \ddots & \ddots & \ddots & \ddots & g_r \\ 0 & \ddots & \ddots & \ddots & \ddots & g_{r+1} \\ \vdots & \ddots & \ddots & \ddots & \ddots & g_{r+2} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \end{bmatrix}, \quad r = \max(m, q).$$

Clearly, the above equation could be solved for the $\{g_i\}$ using the following recursion:

$$\text{(3.26)} \quad zu_{i+1}(z) = u_i(z) - g_i a_t(z); \quad u_0(z) = b(z) \quad \text{and} \quad g_i = u_i(0)/a_t(0).$$

Since the terms $\tau_j^* z^{2(t-j)}a_t(z)$ for all j, $0 \leqq j \leqq t - 1$, do not have any effect on the coefficients of the $t - j$ lowest power of z in the expression, $b_j(z) - \tau_j a_t(z) + \tau_j^* z^{2(t-j)}a_t(z)$, we can conclude that $\{\tau_j\}_{j=0}^{t-1}$ could be computed via (3.24), and in fact $g_j = \tau_j$, $0 \leqq j \leqq t - 1$. Therefore,

$$Q(z, w) - a_t(z)a_t^*(w)\left[\sum_{j=0}^{t-1} [g_j^* z^{t-j} + g_j(w^*)^{t-j}](zw^*)^j[1 - (zw^*)^{t-j}]/(1 - zw^*)\right]$$
$$= (zw^*)^t Q_t^\tau(z, w) = (zw^*)^t Q_{2t}^c(z, w).$$

Hence

$$Q_t^\tau(z, w) = Q_{2t}^c(z, w) = \frac{a_t(z)b_t^*(w) + b_t(z)a_t^*(w)}{1 - zw^*}. \qquad \square$$

Thus (3.23) and (3.24) provide an efficient way of computing the block Schur complement Q_{2t}^c. It is clear that $\{Q_j^\tau\}_{j=1}^{t-1}$ does not represent any Schur complement related to the problem at hand. However, the last member of this finite sequence, $Q_t^\tau(z, w)$, represents the desired block Schur complement and allows us to continue with the factorization procedure. Lemma 3.2 suggests the following block reduction step, namely,

$$(3.27)\qquad \begin{aligned} Q &= A_L \begin{bmatrix} 0 & U^* \\ U & \hat{G}_0 \\ 0 & \hat{G}_1 \\ 0 & \hat{G}_2 \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} 0 & U^* \\ U & \hat{G}_0 \end{bmatrix}^{-1} \begin{bmatrix} 0 & U^* & 0 & 0 & \cdots & \cdots \\ U & \hat{G}_0 & \hat{G}_1^* & \hat{G}_2^* & \cdots & \cdots \end{bmatrix} A_L^* + Q_{2t}^c \\ &= A_L \begin{bmatrix} U^* & 0 \\ \hat{G}_0 & I \\ \hat{G}_1 & 0 \\ \vdots & \vdots \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} 0 & I \\ I & -\hat{G}_0 \end{bmatrix} \begin{bmatrix} U & \hat{G}_0 & \hat{G}_1^* & \hat{G}_2^* & \cdots & \cdots \\ 0 & I & 0 & 0 & \cdots & \cdots \end{bmatrix} A_L^* + Q_{2t}^c. \end{aligned}$$

It is clear from Theorem 3.3 and Lemma 3.1 that

$$U^* = \begin{bmatrix} g_0 & 0 & \cdots & 0 \\ \vdots & g_0 & & \cdots \\ \vdots & & & 0 \\ g_{t-1} & \cdots & \cdots & g_0 \end{bmatrix}_{t\times t}$$

and

$$\hat{G}_0 = \begin{bmatrix} 2\,\mathrm{Re}\,(\hat{g}_0) & \hat{g}_1^* & \cdots & \hat{g}_{t-1}^* \\ \hat{g}_1 & 2\,\mathrm{Re}\,(\hat{g}_0) & & \vdots \\ \vdots & & & \hat{g}_1^* \\ \hat{g}_{t-1} & \cdots & \hat{g}_1 & 2\,\mathrm{Re}\,(\hat{g}_0) \end{bmatrix}_{t\times t}.$$

The generating function of the first t columns of the block triangular factor

$$A_L \begin{bmatrix} U & \hat{G}_0^* & \vdots & \vdots & \vdots & \vdots \\ 0 & I & 0 & \cdots & \cdots & \cdots \end{bmatrix}^* \quad \text{is} \quad a_t(z)\left[\sum_{k=0}^{t-1} (zw^*)^{t-1-k}[g(z) + z^{k+1}\tilde{\gamma}_k(z)]\right],$$

where

$$(3.28)\qquad \tilde{\gamma}_k(z) = \sum_{j=0}^{2t-k-1} g_{2t-j-k-1} z^j; \qquad 0 \leqq k \leqq t-1.$$

It is clear that $\{\tilde{\gamma}_k(z)\}_{k=0}^{t-1}$ can be generated via the backward recursion

$$(3.29)\qquad \begin{cases} \tilde{\gamma}_{k-1}(z) = z\tilde{\gamma}_k(z) + g_{2t-k}, \\ \tilde{\gamma}_{t-1}(z) = \sum_{j=0}^{t} g_{t-j} z^j. \end{cases}$$

Recalling (3.26) we know that $\{g_i\}$ can be computed via

$$zu_{i+1}(z) = u_i(z) - g_i a_t(z); \quad u_0(z) = b(z)$$

and

$$g_i = u_i(0)/a_t(0).$$

Hence the following recursions can be used to compute $\{a_t(z)\tilde{\gamma}_k(z)\}_{k=0}^{t-1}$:

$$(3.30)\qquad zu_{i+1}(z) = u_i(z) - g_i a_t(z)$$

and

$$u_{i+1}(z) = zu_i(z) + g_i a_t(z), \tag{3.31}$$

where

$$g_i = u_i(0)/a_t(0) \quad u_0(z) = b(z) \quad \text{and} \quad u_0(z) = 0. \tag{3.32}$$

It is clear that

$$a_t(z)\tilde{\gamma}_k(z) = \vartheta_{2t-k}(z); \quad 0 \leqq k \leqq t-1. \tag{3.33}$$

Since computing $\{\hat{g}_j\}_{j=0}^{t-1}$ from $\{g_j\}_0^{2t-1}$ is trivial (see Lemma 3.1), computation of $-\hat{G}_0$ for the block diagonal

$$\begin{bmatrix} 0 & I \\ I & -\hat{G}_0 \end{bmatrix}$$

is trivial, also.

3.3. Polynomial recursions and triangular factors. Let us first express the factorization formulas (2.3) and (2.4) (see § 2) in generating function terms,

$$\bar{M}_{j+1}(z) = \bar{M}_j(z, w) - \bar{M}_j(z, 0)\bar{M}_j^{-1}(0, 0)\bar{M}_j(0, w), \tag{3.34}$$

$$d_j = \bar{M}_j(0, 0), \quad \bar{l}_j(z) = \bar{M}_j(z, 0)d_j^{-1} \quad \text{when} \quad d_j \neq 0, \tag{3.35}$$

where

$$(zw^*)^j\bar{M}_j(z, w) = M_j(z, w) = [1 \quad z \quad z^2 \quad \cdots \quad \cdots]M_j[1 \quad w \quad w^2 \quad \cdots \quad \cdots]^*,$$

and

$$z^j\bar{l}_j(z) = l_j(z) = [1 \quad z \quad z^2 \quad \cdots \quad \cdots]\mathbf{l}_j.$$

The following cases represent all the possibilities to be encountered at any step j.

Case 1(a). $\alpha_j(0) = 0$, $|\alpha_j(0)| \neq |\beta_j(0)|$.

This is a strongly regular step (the Schur complement of the (1, 1) entry can be computed), for which the recursion is

$$\begin{cases} z\alpha_{j+1}(z) = \alpha_j(z), \\ \beta_{j+1}(z) = \beta_j(z). \end{cases} \tag{3.36}$$

This implies that

$$Q_j(z, w) = -\beta_j(z)\beta_j^*(w) + (zw^*)Q_{j+1}(z, w),$$

$$d_i = -|\beta_j(0)|^2\left[(-1)^{\delta_j}\prod_{i=0}^{j}\frac{1}{1-|k_i|^2}\right]; \quad |k_i| \neq \infty,$$

and

$$l_i(z) = \beta_{j+1}(z)/\beta_{j+1}(0),$$

where δ_j is the number of times Case 1(a) has occurred before step j.

Case 1(b). $\alpha_j(0) \neq 0$, $|\alpha_j(0)| \neq |\beta_j(0)|$.

This is also a strongly regular step, for which the recursion is

$$\begin{cases} \alpha_{j+1}(z) = \alpha_j(z) - k_j\beta_j(z), \\ z\beta_{j+1}(z) = -k_j^*\alpha_j(z) + \beta_j(z), \end{cases} \tag{3.37}$$

where $k_j = \beta_j^*(0)/\alpha_j^*(0)$. This implies that

$$Q_j(z, w) = \frac{1}{1 - |k_j|^2}\,\alpha_{j+1}(z)\alpha_{j+1}^*(w) + \frac{1}{1 - |k_j|^2}\,zw^*Q_{j+1}(z, w),$$

$$d_i = |\alpha_j(0) - k_j\beta_j(0)|^2\left[(-1)^{\delta_j}\prod_{i=0}^{j}\frac{1}{1 - |k_i|^2}\right]; \quad |k_i| \neq \infty,$$

and

$$l_i(z) = \alpha_{j+1}(z)/\alpha_{j+1}(0),$$

where δ_j is the number of times Case 1(a) has occurred before step j.

Case 2. $\alpha_j(z) = e^{j\phi}\beta_j(z)$ for some $\phi \in [0, 2\pi]$.

Since $Q_j(z, w) = 0$ the factorization is trivial. We simply pick

$$d_j = 0 \quad \text{and} \quad l_j(z) = z^j. \tag{3.38}$$

In fact, at this step there is no need to proceed any further.

Case 3. $\alpha_j(0) = \beta_j(0) = 0$.

Since $Q_j(z, 0) = 0 = Q_j(0, w)$, but $Q_j(z, w) \neq 0$, we proceed (as described in § 2) simply by shifting the matrix Q_j up by one row and one column. Thus we assign

$$zw^*Q_{j+1}(z, w) = Q_j(z, w). \tag{3.39}$$

From (3.39) we get the recursions

$$\begin{cases} z\alpha_{j+1}(z) = \alpha_j(z), \\ z\beta_{j+1}(z) = \beta_j(z), \end{cases} \tag{3.40}$$

which implies that

$$d_j = 0 \quad \text{and} \quad l_j(z) = z^j. \tag{3.41}$$

Case 4. $|\alpha_j(0)| = \beta_j(0)| \neq 0$, $\alpha_j(z) \neq e^{j\phi}\beta_j(z)$ for all $\phi \in [0, 2\pi]$.

In this case we need to find the block factorization of the kind shown in (3.27). Define, following (3.10) (we ignore the factor $\sqrt{(2)}$ here),

$$\begin{cases} a_j(z) = \alpha_j(z) - \dfrac{\beta_j^*(0)}{\alpha_j^*(0)}\,\beta_j(z), \\[2ex] b_j(z) = \alpha_j(z) + \dfrac{\beta_j^*(0)}{\alpha_j^*(0)}\,\beta_j(z). \end{cases} \tag{3.42}$$

Now if t is the smallest positive integer such that $\lim_{z\to 0} z^{-t}a_j(z) \neq 0$, then by Theorem 3.3 the first nonsingular leading principal submatrix of Q_j must be $2t \times 2t$. Then Lemma 3.3 suggests the recursions

$$\begin{cases} za_{j+1}(z) = a_j(z), \\ zb_{j+1}(z) = b_j(z) - \tau_j z^{-t}a_j(z) + \tau_j^* z^t a_j(z), \end{cases} \tag{3.43}$$

where

$$\tau_j = \lim_{z\to 0} z^t b_j(z)/a_j(z).$$

This implies that the block diagonal factor D_j ($2t \times 2t$ matrix) is

$$D_j = (-1)^{\delta_j}\prod_{i=0}^{j-1}\frac{1}{1 - |k_i|^2}\left[\begin{array}{c|c} O & g_{0,j}I_t \\ g_{0,j}I_t & -\hat{G}_{0,j} \end{array}\right]\left|\lim_{z\to 0} z^{-t}a_j(z)\right|^2; \quad |k_i| \neq \infty, \tag{3.44}$$

where $\hat{G}_{0,j}$ is the $t \times t$ Hermitian Toeplitz matrix whose first row is $[2 \operatorname{Re}(\hat{g}_{0,j})\hat{g}^*_{1,j} \cdots \hat{g}^*_{t-1,j}]$, with

$$\hat{g}_{i,j} = \begin{cases} g_{t,j}; & i = 0, \\ g_{t+i,j} + g^*_{t-i,j}; & 1 \leqq i \leqq t, \\ g_{t+i}; & i \geqq t+1, \end{cases}$$

and δ_j is the number of times Case 1(a) has occurred before step j. The $\{g_{i,j}\}$ can be computed via

$$\text{(3.45)} \qquad \begin{cases} zu_{i+1,j}(z) = u_{i,j}(z) - g_{i,j}z^{-t}a_j(z); \quad u_{0,j}(z) = b_j(z) \quad \text{and} \\ g_{i,j} = \lim\limits_{z \to 0} u_{i,j}(z)/z^{-t}a_j(z). \end{cases}$$

The first t columns of the triangular (block) factor have the generating function

$$\text{(3.46)} \qquad \hat{L}_j(z, w) = \frac{[\sum_{k=0}^{t-1} (zw^*)^{t-1-k}[b_j(z) + z^{k+1}a_{j+t}(z)\tilde{\gamma}_{k,j}(z)]]}{|g_{0,j}|\,|a_{j+t}(0)|},$$

where

$$\text{(3.47)} \qquad \tilde{\gamma}_{k,j}(z) = \sum_{i=0}^{2t-k-1} g^*_{2t-i-k-1,j}z^i, \quad 0 \leqq k \leqq t-1,$$

$$\text{(3.48)} \qquad a_{j+t}(z) = z^{-t}a_j(t),$$

and the $\{a_{j+t}(z)\tilde{\gamma}_{k,j}(z)\}$ are computed via

$$\text{(3.49)} \qquad \vartheta_{i+1,j}(z) = z\vartheta_{i,j}(z) + g^*_{i,j}a_{j+t}(z); \quad \vartheta_{0,j}(z) = 0.$$

The last t columns of the block triangular factor have the generating function

$$\text{(3.50)} \qquad \tilde{L}_j(z, w) = a_{j+t}(z)\left[\sum_{i=0}^{t-1} (zw^*)^{i+t}\right].$$

Once the singularity is overcome we will use recursions on $\alpha_j(\cdot)$ and $\beta_j(\cdot)$ rather than the recursions on $a_j(\cdot)$ and $b_j(\cdot)$.

It is important to note that at any step we just need to compute a linear update of two polynomials. This is irrespective of the presence of a singularity. Hence the triangular (modified triangular) factorization of the QT matrices can be computed in $O(n^2)$ computations, using the recursions of this subsection.

Example. Consider the following Toeplitz matrix T_4,

$$T_4 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

whose semi-infinite extension $T_{4,\infty}$ has generating function

$$\text{(3.51)} \qquad T_{4,\infty}(z, w) = \frac{t(z) + t^*(w)}{1 - zw^*}, \quad t(z) = 1 + z.$$

We can rewrite (3.51) as

$$T_{4,\infty}(z, w) = \frac{\alpha(z)\alpha^*(w) - \beta(z)\beta^*(w)}{1 - zw^*}, \quad \alpha(z) = 1 + z \quad \text{and} \quad \beta(z) = z.$$

We initiate the recursions with

$$\alpha_0(z) = 1 + z \quad \text{and} \quad \beta_0(z) = z.$$

It is clear that $\alpha_0(z)$ and $\beta_0(z)$ put us in Case 1(b), so we get $k_0 = 0$. Therefore,

$$\alpha_1(z) = \alpha_0(z) = 1 + z, \qquad \beta_1(z) = \beta_0(z)/z = 1,$$

which implies

$$\mathbf{l}_0(z) = \alpha_1(z) = \alpha_1(z) = 1 + z \quad \text{and} \quad d_0 = 1.$$

Since $|\alpha_1(0)| = |\beta_1(0)|$, but $\alpha_1(z) \neq e^{j\phi}\beta_1(z)$ for all $\phi \in [0, 2\pi]$, the polynomials $\alpha_1(z)$ and $\beta_1(z)$ must correspond to Case 4. We define

$$\sqrt{2}a_1(z) = \alpha_1(z) - \beta_1(z) = z,$$

$$\sqrt{2}b_1(z) = \alpha_1(z) + \beta_1(z) = 2 + z.$$

Since unity is the smallest positive integer such that $z^{-t}a_1(z)|_{z=0} \neq 0$, we assign $t = 1$. Then $a_{t+1}(z) = a_2(z) = 1/\sqrt{2}$ and the 2×2 block diagonal entry D_1 must be

$$D_1 = (-1)^{\delta_0}\frac{1}{(1 - |k_0|^2)}\begin{bmatrix} 0 & g^*_{0,1} \\ g_{0,1} & -\hat{G}_{0,1}\end{bmatrix}|a_2(0)|^2.$$

Since $g_{01} = 2$, $\hat{G}_{01} = \hat{g}_{11} + \hat{g}^*_{11} = 2$, $S_0 = 0$, $k_0 = 0$, and $a_2(0) = 1/\sqrt{2}$, the block diagonal factor is

$$D_1 = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}.$$

Now since $t - 1$, $\hat{L}_1$ and $\tilde{L}_1$ are one column each and

$$\hat{L}_1(z, w) = \frac{[(zw^*)^0[b_1(z) + za_2(z)\tilde{\gamma}_{0,1}(z)]]}{g_{0,1}a_2(0)},$$

where

$$\tilde{\gamma}_{0,1}(z) = g^*_{1,1} + g^*_{0,1}z = 1 + 2z$$

and

$$\tilde{L}_1(z, w) = zw^*a_2(z) = zw^*.$$

Then the generating function $L_1(z, w)$ of the triangular (block) factor must be $\hat{L}_1(z, w) + \tilde{L}_1(z, w)$,

$$L_1(z, w) = 1 + z + z^2 + zw^*.$$

The polynomial $b_2(z)$ is computed as

$$zb_2(z) = b_1(z) - \tau_1 z^{-1}a_1(z) + \tau_1^* za_1(z),$$

where

$$\tau_1 = \lim_{z \to 0}\frac{zb_1(z)}{a_1(z)} = b_1(0) = \sqrt{2}.$$

Then

$$b_2(z) = (1 + 2z)/\sqrt{2}$$

so that

$$\alpha_2(z) = [a_2(z) + b_2(z)]/\sqrt{2} = 1 + z, \qquad \beta_2(z) = [b_2(z) - a_2(z)]/\sqrt{2} = z.$$

Since $\alpha_1(z) = \alpha_2(z)$ and $\beta_1(z) = \beta_2(z)$, it is clear that even the polynomials $\alpha_2(z)$ and $\beta_2(z)$ correspond to Case 4, and we get the same diagonal and triangular block factors. We can proceed in this way to get the factorization

$$T_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

after proper truncation of the factors of $T_{4,\infty}$.

4. Inertia of the block diagonal factor. Computing the inertia of Toeplitz and QT matrices has applications in the problems of root distribution of polynomials with respect to the unit circle. If the given matrix is strongly regular, the inertia can be determined from the diagonal factor D by inspecting the signs of the nonzero entries of D. In this section we show that the situation is not really different even when the given matrix is not strongly regular. The reason is that the signature of a block diagonal entry is zero, as we will now prove.

The diagonal factor D_j in (3.44) has the same inertia as the matrix

$$\left[\begin{array}{c|c} O & I_t \\ \hline I_t & -\hat{G}_{0,j}/g_{0,j} \end{array}\right].$$

Since $\hat{G}_{0,j}$ is Hermitian,

$$(4.1) \qquad \left[\begin{array}{c|c} I_t & O \\ \hline \hat{G}_{0,j}/2g_{0,j} & I_t \end{array}\right] \left[\begin{array}{c|c} O & I_t \\ \hline I_t & -\hat{G}_{0,j}/g_{0,j} \end{array}\right] \left[\begin{array}{c|c} I_t & \hat{G}^*_{0,j}/2g_{0,j} \\ \hline O & I_t \end{array}\right] = \left[\begin{array}{c|c} O & I_t \\ \hline I_t & O \end{array}\right];$$

which implies that D_j must have an equal number of positive and negative eigenvalues. In fact, the above result also enables a simple proof of Iohvidov's rules (see Iohvidov [Ioh82]) for determination of the nullity and inertia of Hermitian Toeplitz matrices with arbitrary rank profile. Such an alternative proof will be published elsewhere.

5. Inversion of Toeplitz and quasi-Toeplitz matrices: Block generating function approach. If possible, let the Hermitian Toeplitz matrix T_{n-1} of (3.2) be nonsingular. We will define the $2n \times 2n$ Toeplitz block matrix

$$(5.1) \qquad T_b = \left[\begin{array}{c|c} T_{n-1} & I \\ I & O \end{array}\right]_{2n \times 2n}.$$

T_{n-1} is the first $n \times n$ principal submatrix of T_b, and the corresponding block Schur complement is $-T_{n-1}^{-1}$. We define the following semi-infinite matrix

$$(5.2) \qquad T_{b,\infty} = \left[\begin{array}{c|c} T_{n-1,\infty} & I_\infty \\ I_\infty & O_\infty \end{array}\right],$$

where $T_{n-1,\infty}$ is the semi-infinite extension of T_{n-1}, I_∞ is a semi-infinite identity matrix, and O_∞ is a semi-infinite zero matrix. It is useful to associate a semi-infinite block structured matrix $M = \left[\begin{smallmatrix} M_{11} & | & M_{12} \\ M_{21} & | & M_{22} \end{smallmatrix}\right]$ (with structured semi-infinite blocks M_{ij}, $1 \leqq i, j \leqq 2$) with a block "generating" function

$$(5.3)\quad \begin{aligned} M(z, w) &= \begin{bmatrix} 1 & z & z^2 & \cdots & | & O & \cdots & \cdots & O \\ O & \cdots & \cdots & O & | & 1 & z & z^2 & \cdots \end{bmatrix} \begin{bmatrix} M_{11} & | & M_{12} \\ M_{21} & | & M_{22} \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} 1 & w & w^2 & \cdots & | & O & \cdots & \cdots & O \\ O & \cdots & \cdots & O & | & 1 & w & w^2 & \cdots \end{bmatrix}^* . \end{aligned}$$

Then the block generating function (which is obtained by replacing each of the semi-infinite Toeplitz blocks by their generating functions)

$$(5.4)\qquad T_{b,\infty}(z, w) = \begin{bmatrix} T_{n-1,\infty}(z, w) & | & \dfrac{1}{1 - zw^*} \\ \dfrac{1}{1 - zw^*} & | & O \end{bmatrix}_{2 \times 2},$$

where $T_{n-1,\infty}(z, w)$ is as in (3.2).

For simplicity assume $t_0 = 1$ for now. Then notice that

$$(5.5)\qquad T_{b,\infty}(z, w) = \frac{G_b(z) J G_b^*(w)}{1 - zw^*},$$

where

$$(5.6)\qquad J = \begin{bmatrix} 1 & | & O \\ O & | & -1 \end{bmatrix} \quad \text{and} \quad G_b(z) = \begin{bmatrix} t(z) + \frac{1}{2} & | & t(z) - \frac{1}{2} \\ 1 & | & 1 \end{bmatrix}.$$

Next we will define a matrix Q_b to be BQT if its block generating function $Q_b(z, w)$ can be represented in the form

$$(5.7)\quad Q_b(z, w) = \frac{1}{1 - zw^*} \begin{bmatrix} \alpha(z) & | & \beta(z) \\ \gamma(z) & | & \delta(z) \end{bmatrix} \begin{bmatrix} 1 & | & O \\ O & | & -1 \end{bmatrix} \begin{bmatrix} \alpha(w) & | & \beta(w) \\ \gamma(w) & | & \delta(w) \end{bmatrix}^*,$$

where $\alpha(z)$, $\beta(z)$, $\gamma(z)$, and $\delta(z)$ are all polynomials (functions) in z.

THEOREM 5.1. *The Schur complement of the* 1×1 *leading principal submatrix of a* BQT *matrix is also* BQT *provided the* 1×1 *entry is nonzero.*

Proof. The generating function $Q_b^c(z, w)$ of the Schur complement is given by

$$(5.8)\quad \begin{aligned} &\begin{bmatrix} z & | & O \\ O & | & 1 \end{bmatrix} Q_b^c(z, w) [w^* \quad | \quad O] \\ &\quad = Q_b(z, w) - Q_b(z, 0)[1 \quad 0]^* ([1 \quad 0] Q_b(0, 0) [1 \quad 0]^*)^{-1} [1 \quad 0] Q_b(0, w), \end{aligned}$$

where $Q_b(z, w)$ in (5.8) is the block generating function of BQT matrix Q_b in (5.7). Assuming that $Q_b(0, 0) \neq 0$ we obtain, after some algebra,

$$(5.9)\quad Q_b^c(z, w) = \frac{\begin{bmatrix} \alpha_1(z) & | & \beta_1(z) \\ z\gamma_1(z) & | & \delta_1(z) \end{bmatrix} \begin{bmatrix} 1 & | & O \\ O & | & -1 \end{bmatrix} \begin{bmatrix} \alpha_1(w) & | & \beta_1(w) \\ w\gamma_1(w) & | & \delta_1(w) \end{bmatrix}^*}{(|\alpha(0)|^2 - |\beta(0)|^2)(1 - zw^*)},$$

where

$$(5.10)\qquad \begin{aligned} \alpha_1(z) &= \alpha(z)\alpha^*(0) - \beta(z)\beta^*(0), & \beta_1(z) &= [\alpha(z)\beta(0) - \beta(z)\alpha(0)]/z,\\ \gamma_1(z) &= \gamma(z)\alpha^*(0) - \delta(z)\beta^*(0), \quad \text{and} & \delta_1(z) &= [\gamma(z)\beta(0) - \delta(z)\alpha(0)]. \end{aligned}$$

It is clear that $\beta_1(z)$ is a polynomial since the numerator in (5.10) has a root at the origin. This completes the proof of the theorem, since $Q_b^c(z, w)$ is clearly BQT. □

It is clear that n steps of Schur reduction on the block generating function $T_{b,\infty}(z, w)$ would lead directly to the generating function of T_{n-1}^{-1} via the computed polynomials (functions) $\gamma_{n-1}(z)$ and $\delta_{n-1}(z)$ starting with $\gamma_0(z) = 1$ and $\delta_0(z) = 1$ as in (5.7).

It was shown in [Pal90, Appendix A, § A.2, eq. (A.5)] that for an arbitrary nonsingular QT matrix Q_{n-1} it is $Q_{n-1}^{-\#}$ and not Q_{n-1}^{-1} that possesses the same displacement rank.[1] This naturally raises the following question: Is it possible to compute Q_{n-1}^{-1} in a similar fashion? The answer is "yes" and we provide an explanation below.

It is clear that

$$(5.11\text{a})\qquad \begin{aligned} Q_{n-1} &= L_{n-1}\{\underline{\alpha}\}L_{n-1}^*\{\underline{\alpha}\} - L_{n-1}\{\underline{\beta}\}L_{n-1}^*\{\underline{\beta}\}\\ &= L_{n-1}\{\underline{\alpha} - \underline{\beta}\}TL_{n-1}^*\{\underline{\alpha} - \underline{\beta}\}, \end{aligned}$$

where

$$(5.11\text{b})\qquad T = \frac{1}{2}\{L_{n-1}\{\underline{\alpha} + \underline{\beta}\}L_{n-1}^{-1}\{\underline{\alpha} - \underline{\beta}\} + L_{n-1}^*\{\underline{\alpha} + \underline{\beta}\}L_{n-1}^{-*}\{\underline{\alpha} - \underline{\beta}\}\}$$

(assuming invertibility of $L_{n-1}\{\underline{\alpha} - \underline{\beta}\}$; otherwise we will use $-\underline{\beta}$ instead of $\underline{\beta}$) is a nonsingular Hermitian Toeplitz matrix

$$(5.12\text{a})\qquad \begin{aligned} Q_{n-1}^{-\#} &= L_{n-1}^{-1}\{\underline{\alpha} - \underline{\beta}\}T^{-1}L_{n-1}^{-*}\{\underline{\alpha} - \underline{\beta}\}\\ &= L_{n-1}^{-1}\{\underline{\alpha} - \underline{\beta}\}L_{n-1}^*\{\underline{\alpha} - \underline{\beta}\}Q_{n-1}^{-1}L_{n-1}\{\underline{\alpha} - \underline{\beta}\}L_{n-1}^{-*}\{\underline{\alpha} - \underline{\beta}\}, \end{aligned}$$

by the fact that $T^{-1} = T^{-\#}$. Given $\alpha(z)$ and $\beta(z)$ it is possible to find $\gamma_e(z)$ such that

$$(5.12\text{b})\qquad [\alpha(z) - \beta(z)]\gamma_e(z) = 1 + O(z^n).$$

Partition $\gamma_e(z)$ as $\gamma_e(z) = \hat{\gamma}_e(z) + O(z^n)$. Then the choice

$$(5.12\text{c})\qquad \gamma(z) = \delta(z) = \hat{\gamma}_e(z)$$

would be sufficient and the Schur procedure would compute the generators for $Q^{-\#}$ in n steps.

5.1. Admissibility condition. It is clear that construction of the generators for an arbitrary nonsingular Hermitian QT matrix requires a certain amount of preprocessing for obtaining the starting $\gamma(z)$ and $\delta(z)$. This is equivalent to the prefiltering in Lev-Ari and Kailath [LK84] or in Kailath, Bruckstein, and Morgan [KBM86].

If the preprocessing is not desired, we can always possibly construct generators of displacement rank four for the inverse working with a nonminimal generator for the

[1] $Q^\# = \tilde{I}Q'\tilde{I}$ is sometimes called Q-natural; $\tilde{I}$ being the reversed identity matrix and $'$ being the transpose.

original matrix. We can choose

(5.13a)

$$Q_b(z,w)=\frac{\left[\begin{array}{c|c|c|c}\alpha(z) & \beta(z) & 1 & 1\\ O & O & \frac{1}{2} & -\frac{1}{2}\end{array}\right]\left[\begin{array}{c|c|c|c}1 & O & O & O\\ O & -1 & O & O\\ O & O & 1 & O\\ O & O & O & -1\end{array}\right]\left[\begin{array}{c|c}\alpha^*(w) & O\\ \beta^*(w) & O\\ 1 & \frac{1}{2}\\ 1 & -\frac{1}{2}\end{array}\right]}{1-zw^*},$$

which implies that

(5.13b) $$Q_{b,\infty}=\left[\begin{array}{c|c}Q_{n-1,\infty} & I_\infty\\ I_\infty & O_\infty\end{array}\right].$$

Thus, after n steps of Schur complementation we obtain the generators for the inverse. However, if the given $\alpha(z)$ and $\beta(z)$ are such that there exists complex numbers λ and μ satisfying

(5.14a) $$\lambda\alpha(z)-\mu\beta(z)=1,$$

then it is possible to find generators of displacement rank three while working with generators of displacement rank two. This condition (5.14a) is a so-called admissibility condition (see [LK84]). Next choose

(5.14b) $$Q_b(z,w)=\frac{1}{1-zw^*}\left[\begin{array}{c|c}\alpha(z) & \beta(z)\\ \lambda & \mu\end{array}\right]\left[\begin{array}{c|c}1 & O\\ O & -1\end{array}\right]\left[\begin{array}{c|c}\alpha(w) & \beta(w)\\ \lambda & \mu\end{array}\right]^*,$$

which implies that

(5.14c) $$Q_{b,\infty}=\left[\begin{array}{c|c}Q_{n-1,\infty} & I_\infty\\ I_\infty & (\lambda\lambda^*-\mu\mu^*)I_\infty\end{array}\right].$$

As usual, after n steps of Schur complementation we obtain the generators for the inverse. However, depending on the sign of $\lambda\lambda^*-\mu\mu^*$, the generator changes. If $\lambda\lambda^*-\mu\mu^*>0$, the generator must be $[\gamma_n(z)\delta_n(z)(\lambda\lambda^*-\mu\mu^*)^{1/2}]$ with $J=-1\oplus 1\oplus 1$; otherwise, if $\lambda\lambda^*-\mu\mu^*<0$, then the generator must be $[\gamma_n(z)\delta_n(z)(\lambda\lambda^*-\mu\mu^*)^{1/2}]$ with $J=-1\oplus 1\oplus -1$.

It is clear that as long as $|\alpha(0)|^2\neq|\beta(0)|^2$, the above procedure works and the following recursions could be used to update $\alpha(z)$, $\beta(z)$, $\gamma(z)$, and $\delta(z)$.

5.2. Singularities and block Schur complements. The Schur complement $Q_b^c(z,w)$ of the 1×1 leading principal submatrix of a BQT matrix cannot be computed if $[1\quad 0]Q_b(0,0)[1\quad 0]^*=Q(0,0)=0$ (see (5.8)). Since we are interested in inverting Q, the first two cases (see § 3.2) of singularity are not material. The third case, which requires computing a block Schur complement, must be understood in terms of its impact on the block generating function procedure.

5.2.1. Immittance variables and the structure of Q_b. It is convenient to understand singularities in terms of the immittance variables. Motivated by this we rewrite the expression (5.6) using the following so-called immittance variables:

(5.15) $$\left\{\begin{array}{l}\sqrt{2}a(z)=\alpha(z)-(\beta^*(0)/\alpha^*(0))\beta(z),\\ \sqrt{2}b(z)=\alpha(z)+(\beta^*(0)/\alpha^*(0))\beta(z),\\ \sqrt{2}c(z)=\gamma(z)-(\beta^*(0)/\alpha^*(0))\delta(z),\\ \sqrt{2}d(z)=\gamma(z)+(\beta^*(0)/\alpha^*(0))\delta(z).\end{array}\right.$$

We get, say,

$$\begin{aligned} Q_b(z, w) &= \frac{\left[\begin{array}{c|c} \alpha(z) & \beta(z) \\ \gamma(z) & \delta(z) \end{array}\right]\left[\begin{array}{c|c} 1 & O \\ O & -1 \end{array}\right]\left[\begin{array}{c|c} \alpha(w) & \beta(w) \\ \gamma(w) & \delta(w) \end{array}\right]^*}{(1 - zw^*)} \\ &= \frac{\left[\begin{array}{c|c} a(z)b^*(w) + b(z)a^*(w) & a(z)d^*(w) + b(z)c^*(w) \\ a^*(w)d(z) + b^*(w)c(z) & c(z)d^*(w) + d(z)c^*(w) \end{array}\right]}{(1 - zw^*)} \\ &= \frac{\left[\begin{array}{c|c} a(z) & b(z) \\ c(z) & d(z) \end{array}\right]\left[\begin{array}{c|c} O & 1 \\ 1 & O \end{array}\right]\left[\begin{array}{c|c} a(w) & b(w) \\ c(w) & d(w) \end{array}\right]^*}{(1 - zw^*)} \\ &= \left[\begin{array}{c|c} Q(z, w) & Q_X^*(w, z) \\ Q_X(z, w) & Q_l(z, w) \end{array}\right]. \end{aligned} \tag{5.16}$$

Let $a(z) = z^t a_t(z)$ (see Theorem 3.3). Since congruence preserves rank profile and inertia, the BQT form $G_b(z, w)$ defined as, say,

$$\left[\begin{array}{c|c} a_t(z) & O \\ O & c(z) \end{array}\right] G_b(z, w) \left[\begin{array}{c|c} a_t^*(w) & O \\ O & c^*(w) \end{array}\right] = \left[\begin{array}{c|c} Q(z, w) & Q_X^*(w, z) \\ Q_X(z, w) & Q_l(z, w) \end{array}\right], \tag{5.17}$$

must have the same rank profile and inertia as $Q_b(z, w)$. Let $G_b(z, w)$ be partitioned as, say,

$$G_b(z, w) = \left[\begin{array}{c|c} \tilde{G}(z, w) & G_X^*(w, z) \\ G_X(z, w) & G_l(z, w) \end{array}\right]. \tag{5.18}$$

5.2.2. Structures of $\tilde{G}$, G_X, and, G_l. The structure of $\tilde{G}$ has been studied in § 3. Let us first examine the structure of G_X:

$$G_X(z, w) = \frac{x(z)(w^*)^t + g^*(w)}{1 - zw^*}, \tag{5.19a}$$

where

$$x(z) = d(z)/c(z) = \sum_{i=0}^{\infty} x_i. \tag{5.19b}$$

Then

$$\begin{aligned} G_X &= \left[\begin{array}{c|ccc} U & \cdot & \cdot & \cdot \\ O & \cdot & \cdot & \cdot \\ \vdots & \cdot & \widehat{G_X} & \cdot \\ \vdots & \cdot & \cdot & \cdot \\ O & \cdot & \cdot & \cdot \end{array}\right] \\ &= \left[\begin{array}{c|c|cccc} U & L + \widehat{G_{x0}} & \widehat{G_{x1}} & \widehat{G_{x2}} & \cdots & \cdots \\ \hline O & U + \hat{X}_0 & L + \widehat{G_{x0}} & \widehat{G_{x1}} & \ddots & \ddots \\ O & \hat{X}_1 & U + \hat{X}_0 & L + \widehat{G_{x0}} & \ddots & \ddots \\ O & \hat{X}_2 & \hat{X}_1 & \ddots & \ddots & \ddots \\ O & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{array}\right], \end{aligned} \tag{5.20}$$

where U and L are $t \times t$ upper and lower triangular Toeplitz matrices, respectively, and $\widehat{G_X}$ is a semi-infinite Toeplitz matrix with the $t \times t$ partitions shown above. It should be noted that U is as defined in (3.24). L is a $t \times t$ lower triangular Toeplitz matrix whose first column is $[x_0^* \quad x_1^* \quad \cdots \quad \cdots \quad x_{t-1}^*]^*$. The matrices $\{\widehat{G_{xj}}\}$ and $\{\widehat{G_j}\}$, $1 \leqq j \leqq \infty$, are related as follows:

$$\text{(5.21)} \qquad \begin{cases} \hat{G}_1^* = \widehat{G_{x1}}^* + U^*, \\ \hat{G}_j = \widehat{G_{xj}}, \quad 2 \leqq j \leqq \infty. \end{cases}$$

The $2t \times 2t$ block Schur complementation on G_b modifies G_X and produces G_X^c as follows:

$$\left[\begin{array}{cc|cccc} O & O & \cdot & \cdot & \cdot & \cdot \\ O & O & \cdot & \cdot & \cdot & \cdot \\ O & O & \cdot & \cdot & \cdot & \cdot \\ O & O & \cdot & G_X^c & \cdot & \cdot \\ O & O & \cdot & \cdot & \cdot & \cdot \\ O & O & \cdot & \cdot & \cdot & \cdot \\ O & O & \cdot & \cdot & \cdot & \cdot \end{array}\right] = G_X - \begin{bmatrix} U & L + \widehat{G_{x0}} \\ O & U + \hat{X}_0 \\ \text{—} & \text{—} \\ O & \hat{X}_1 \\ \vdots & \vdots \end{bmatrix}$$

$$\text{(5.22)} \qquad \cdot \begin{bmatrix} O & U^* \\ U & \hat{G}_0 \end{bmatrix}^{-1} \left[\begin{array}{cc|cc} O & U^* & O & \cdots \\ U & \hat{G}_0 & \hat{G}_1^* & \cdots \end{array}\right]$$

$$= \left[\begin{array}{cc|cccc} O & O & -U^* & O & O & \cdots \\ \hline O & O & \cdot & \cdot & \cdot & \cdot \\ O & O & \cdot & \cdot & \cdot & \cdot \\ O & O & \cdot & \widehat{G_X} & \cdot & \cdot \\ O & O & \cdot & \cdot & \cdot & \cdot \\ O & O & \cdot & \cdot & \cdot & \cdot \end{array}\right].$$

It can be shown that the generating function $G_X^c(z, w)$ of the above matrix G_X^c is given by

$$\text{(5.23a)} \quad G_X^c(z, w) = \frac{[-\tilde{g}_t(z) + z^t x(z) + z^{t+1}\tilde{g}_t^{\#}(z)] + z^t[\hat{g}_t^*(w) + w^*\tilde{g}_t^{\#*}(w)]}{1 - zw^*},$$

where $x(z)$ is as defined above and

$$\text{(5.23b)} \qquad \begin{cases} \tilde{g}_t(z) = g_0 + g_1 z + \cdots + g_{t-1} z^{t-1}, \\ \hat{g}_t(z) = g_t + g_{t+1} z + \cdots \cdots. \end{cases}$$

Noting that $\hat{g}(z) = \hat{g}_t(z) + z\tilde{g}_t^{\#}(z)$ (see Lemma 3.1) and defining $\hat{x}(z) = -\tilde{g}_t(z) + z^t x(z) + z^{t+1}\tilde{g}_t^{\#*}(w)$, we can rewrite (5.23a) as

$$\text{(5.24)} \qquad G_X^c(z, w) = \frac{\hat{x}(z) + z^t \hat{g}^*(w)}{1 - zw^*},$$

so, clearly, G_X^c is QT. Next we examine the structure of G_I:

$$\text{(5.25)} \qquad G_I(z, w) = \frac{x(z) + x^*(w)}{1 - zw^*},$$

where $x(z)$ has been defined above. Then

$$(5.26)\qquad G_I = \left[\begin{array}{cc|cccc} 2\,\mathrm{Re}\,\{\hat{X}_0\} & \hat{X}_1^* & \hat{X}_2^* & \cdots & \cdots \\ \hat{X}_1 & 2\,\mathrm{Re}\,\{\hat{X}_0\} & \hat{X}_1^* & \ddots & \ddots \\ \hline \hat{X}_2 & \hat{X}_1 & 2\,\mathrm{Re}\,\{\hat{X}_0\} & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{array}\right].$$

The $2t \times 2t$ block Schur complementation on G_b modifies G_I and produces G_I^c as follows:

$$G_I^c = G_I - \left[\begin{array}{cc} U & L + \widehat{G_{x0}} \\ O & U + \hat{X}_0 \\ \hline O & \hat{X}_1 \\ \vdots & \vdots \\ \vdots & \vdots \end{array}\right] \begin{bmatrix} O & U^* \\ U & \hat{G}_0 \end{bmatrix}^{-1}$$

$$(5.27)\qquad \cdot\left[\begin{array}{cc|cccc} U^* & O & O & O & \cdots & \cdots \\ L^* + \widehat{G_{x0}}^* & U^* + \hat{X}_0^* & \hat{G}_1^* & \hat{G}_2^* & \cdots & \cdots \end{array}\right]$$

$$= \left[\begin{array}{cc|ccccc} O & -U^* & O & O & O & \cdots & \\ \hline -U & & \cdot & \cdot & \cdot & \cdot & \cdot \\ O & & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & & \cdot & \cdot & \widehat{G_I} & \cdot & \cdot \\ \vdots & & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right].$$

It can be shown that the generating function $G_I^c(z, w)$ of the above matrix G_I^c is given by

(5.28)

$$G_I^c(z, w) = \frac{[-\tilde{g}_t(z) + z^t x(z) + z^{t+1}\tilde{g}_t^{\#}(z)](w^*)^t + [-\tilde{g}_t^*(w) + (w^*)^t x^*(w) + (w^*)^{t+1}\tilde{g}_t^{\#*}(w)]z^t}{1 - zw^*}$$

$$= \frac{\hat{x}(z)(w^*)^t + \hat{x}^*(w)z^t}{1 - zw^*},$$

where $\tilde{x}(z)$ has been defined before. Thus, clearly, G_I^c is QT. Then the generating function of the $2t \times 2t$ block Schur complement $G_{b_{2t}}^c$ of G_b is given by

$$G_{b_{2t}}^c(z, w) = \left[\begin{array}{c|c} G_{2t}^c(z, w) & G_{X_{2t}}^{c*}(z, w) \\ G_{X_{2t}}^c(z, w) & G_{I_{2t}}^c(z, w) \end{array}\right]$$

$$(5.29)\qquad = \frac{\left[\begin{array}{c|c} \hat{g}(z) + \hat{g}^*(w) & \hat{x}^*(w) + (w^*)^t\hat{g}(z) \\ \hat{x}(z) + z^t\hat{g}^*(w) & \hat{x}(z)(w^*)^t + z^t\hat{x}^*(w) \end{array}\right]}{1 - zw^*}$$

$$= \frac{\left[\begin{array}{c|c} 1 & \hat{g}(z) \\ z^t & \hat{x}(z) \end{array}\right]\left[\begin{array}{c|c} O & 1 \\ 1 & O \end{array}\right]\left[\begin{array}{c|c} 1 & \hat{g}(w) \\ w^t & \hat{x}(w) \end{array}\right]^*}{(1 - zw^*)}.$$

Thus the $2t \times 2t$ block Schur complement G_I^c of G_I must be BQT, which implies that by congruence the $2t \times 2t$ block Schur complement Q_I^c of Q_I must also be BQT, that is, say,

$$
\begin{aligned}
Q_I^c(z, w) &= \frac{\left[\begin{array}{c|c} a_t(z) & a_t(z)\hat{g}(z) \\ z^t c(z) & c(z)\hat{x}(z) \end{array}\right]\left[\begin{array}{c|c} O & 1 \\ 1 & O \end{array}\right]\left[\begin{array}{c|c} a_t(w) & a_t(w)\hat{g}(w) \\ w^t c(w) & c(w)\hat{x}(w) \end{array}\right]^*}{(1 - zw^*)} \\
&= \frac{\left[\begin{array}{c|c} a_t(z) & b_t(z) \\ c_t(z) & d_t(z) \end{array}\right]\left[\begin{array}{c|c} O & 1 \\ 1 & O \end{array}\right]\left[\begin{array}{c|c} a_t(w) & b_t(w) \\ c_t(w) & d_t(w) \end{array}\right]^*}{(1 - zw^*)}.
\end{aligned}
\tag{5.30a}
$$

Then

$$
b_t(z) = a_t(z)\hat{g}(z), \quad c_t(z) = z^t c(z) \quad \text{and} \quad d_t(z) = c(z)\hat{x}(z) = z^{-t} c_t(z). \tag{5.30b}
$$

It is clear that $b_t(z)$ is as defined by (3.23) and (3.24). Since it is possible to compute $b_t(z)$ using a linear scalar recursion (see (3.24)), it may be possible to obtain one such recursion for computing $d_t(z)$. Next we show that it is indeed so.

5.2.3. Recursions for updating *a*(*z*), *b*(*z*), *c*(*z*), and *d*(*z*). Since $g_j = \tau_j$, $0 \leqq j \leqq t - 1$, $\tilde{g}_t(z) = \tau_0 + \tau_1 z + \cdots + \tau_{t-1} z^{t-1}$. Thus it is clear from the definition of $\hat{x}(z)$ that the linear recursion

$$
\begin{cases}
x_0(z) = z^t x(z), \\
x_{j+1}(z) = x_j(z) - \tau_j + \tau_j^* z^{2(t-j)}, \\
x_{t-1}(z) = \hat{x}(z),
\end{cases}
\tag{5.31a}
$$

recursively computes $\hat{x}(z)$. So the following linear recursion

$$
\begin{cases}
c_0(z) = c(z) \quad \text{and} \quad d_0(z) = d(z), \\
z^{-1} c_{j+1}(z) = c_j(z), \\
z^{-1} d_{j+1}(z) = d_j(z) - \tau_j z^{-(t-j)} c_j(z) + \tau_j^* z^{(t-j)} c_j(z),
\end{cases}
\tag{5.31b}
$$

recursively computes $c_t(z)$ and $d_t(z)$ using the $\{\tau_j\}$ in (3.43). Then we can combine the recursions (3.43) with the above recursion (5.31b) and write

$$
\begin{aligned}
&\left[\begin{array}{c|c} z & O \\ O & z^{-1} \end{array}\right]\left[\begin{array}{c|c} a_{j+1}(z) & b_{j+1}(z) \\ c_{j+1}(z) & d_{j+1}(z) \end{array}\right] \\
&\qquad = \left[\begin{array}{c|c} a_j(z) & b_j(z) \\ c_j(z) & d_j(z) \end{array}\right]\left[\begin{array}{c|c} 1 & -\tau_j z^{-(t-j)} + \tau_j^* z^{t-j} \\ O & 1 \end{array}\right].
\end{aligned}
\tag{5.32}
$$

5.3. Polynomial recursions for Schur complementation. Using (5.10) we can rewrite (5.9) as follows:

$$
\begin{aligned}
Q_b^c(z, w) &= \frac{\left[\begin{array}{c|c} \alpha_1(z) & \beta_1(z) \\ z\gamma_1(z) & \delta_1(z) \end{array}\right]\left[\begin{array}{c|c} 1 & O \\ O & -1 \end{array}\right]\left[\begin{array}{c|c} \alpha_1(w) & \beta_1(w) \\ w\gamma_1(w) & \delta_1(w) \end{array}\right]^*}{(|\alpha(0)|^2 - |\beta(0)|^2)(1 - zw^*)} \\
&= \frac{\left[\begin{array}{c|c} z & O \\ O & 1 \end{array}\right]\left[\begin{array}{c|c} \alpha(z) & \beta(z) \\ \gamma(z) & \delta(z) \end{array}\right]\Theta(z) J \Theta^*(w)\left[\begin{array}{c|c} \alpha(w) & \beta(w) \\ \gamma(w) & \delta(w) \end{array}\right]^*}{1 - zw^*},
\end{aligned}
\tag{5.33}
$$

where $J = 1 \oplus -1$ and the 2×2 matrix $\Theta(z)$ assumes three different forms which are discussed in the following.

$\Theta(z)$ for a strongly regular step. It is clear that a strongly regular step of Schur complementation is translated into a univariate map $G(z) \to G(z)\Theta(z)$ of form

$$\begin{bmatrix} z & | & O \\ O & | & 1 \end{bmatrix} G_{j+1}(z) = G_j(z)\Theta_j(z). \tag{5.34}$$

Let us define the partitions

$$G_j(z) = \begin{bmatrix} \alpha_j(z) & | & \beta_j(z) \\ \gamma_j(z) & | & \delta_j(z) \end{bmatrix}. \tag{5.35}$$

It should be noted that the transformation $\Theta_j(z)$ is completely determined by $\alpha_j(z)$ and $\beta_j(z)$.

Case 1(a). $\alpha_j(0) = 0$, $|\alpha_j(0)| \neq |\beta_j(0)|$.

This is a strongly regular step (the Schur complement of the (1, 1) entry can be computed), for which the recursion is

$$\begin{cases} z\alpha_{j+1}(z) = \alpha_j(z), \\ \beta_{j+1}(z) = \beta_j(z). \end{cases} \tag{5.36a}$$

This implies that

$$\Theta_j(z) = \begin{bmatrix} 1 & | & O \\ O & | & z \end{bmatrix}. \tag{5.36b}$$

Case 1(b). $\alpha_j(0) \neq 0$, $|\alpha_j(0)| \neq |\beta_j(0)|$.

This is also a strongly regular step, for which the recursion is

$$\begin{cases} \alpha_{j+1}(z) = \alpha_j(z) - k_j\beta_j(z), \\ z\beta_{j+1}(z) = -k_j^*\alpha_j(z) + \beta_j(z), \end{cases} \tag{5.37a}$$

where $k_j = \alpha_j(0)/\beta_j(0)$.

This implies that

$$\Theta_j(z) = \frac{1}{(1 - k_jk_j^*)^{1/2}} \begin{bmatrix} 1 & | & -k_j \\ -k_j^* & | & 1 \end{bmatrix} \begin{bmatrix} z & | & O \\ O & | & 1 \end{bmatrix} \tag{5.37b}$$

if $|k_j| < 1$; otherwise, if $|k_j| > 1$, then

$$\Theta_j = \frac{1}{(k_jk_j^* - 1)^{1/2}} \begin{bmatrix} 1 & | & -k_j \\ -k_j^* & | & 1 \end{bmatrix} \begin{bmatrix} z & | & O \\ O & | & 1 \end{bmatrix}. \tag{5.37c}$$

$\Theta(z)$ for a nonstrongly regular step.

Case 2. $\alpha_j(z) = e^{j\phi}\beta_j(z)$ for some $\phi \in [0, 2\pi]$.

Since $Q_j(z, w) = 0$, all the following principal submatrices of Q are singular. Thus at this step there is no need to proceed any further.

Case 3. $\alpha_j(0) = \beta_j(0) = 0$.

Since $Q_j(z, 0) = 0 = Q_j(0, w)$, Q_j and all the following principal submatrices are singular and so there is no need to proceed any further.

Case A. $|\alpha_j(0)| = \beta_j(0)| \neq 0$, $\alpha_j(z) \neq e^{j\phi}\beta_j(z)$ for all $\phi \in [0, 2\pi]$.

In this case we need to find a $2t_j \times 2t_j$ block Schur complement provided

$$a_j(z) = \alpha_j(z) - \frac{\alpha_j(0)}{\beta_j(0)}\beta_j(z)$$

is not identically zero and has t_j repeated zero roots. At this point we need to make repeated (t_j times) use of the linear recursion

$$(5.38\text{a})\qquad \begin{cases} za_{j+1}(z) = a_j(z), \\ zb_{j+1}(z) = b_j(z) - \tau_j z^{-t_j} a_j(z) + \tau_j^* z^{t_j} a_j(z), \\ z^{-1}c_{j+1}(z) = c_j(z), \\ z^{-1}d_{j+1}(z) = d_j(z) - \tau_j z^{-t_j} c_j(z) + \tau_j^* z^{t_j} c_j(z), \end{cases}$$

where

$$\tau_j = \lim_{z \to 0} z^t b_j(z)/a_j(z)$$

to obtain the generators for the $2t_j \times 2t_j$ block Schur complement. The above recursion (5.38a) can be rewritten in terms of $\alpha_j(z)$, $\beta_j(z)$, $\gamma_j(z)$, and $\delta_j(z)$ as follows:

$$(5.38\text{b})\qquad \begin{cases} z\alpha_{j+1}(z) = (1 + \frac{\tau_j^*}{2} z^{t_j} - \frac{\tau_j}{2} z^{-t_j})\alpha_j(z) - \frac{k_j}{2}(\tau_j^* z^{t_j} - \tau_j z^{-t_j})\beta_j(z), \\ z\beta_{j+1}(z) = \frac{k_j^*}{2}(\tau_j^* z^{t_j} - \tau_j z^{-t_j})\alpha_j(z) + (1 - \frac{\tau_j^*}{2} z^{t_j} + \frac{\tau_j}{2} z^{-t_j})\beta_j(z), \\ z^{-1}\gamma_{j+1}(z) = (1 + \frac{\tau_j^*}{2} z^{t_j} - \frac{\tau_j}{2} z^{-t_j})\gamma_j(z) - \frac{k_j}{2}(\tau_j^* z^{t_j} - \tau_j z^{-t_j})\delta_j(z), \\ z^{-1}\delta_{j+1}(z) = \frac{k_j^*}{2}(\tau_j^* z^{t_j} - \tau_j z^{-t_j})\gamma_j(z) + (1 - \frac{\tau_j^*}{2} z^{t_j} + \frac{\tau_j}{2} z^{-t_j})\delta_j(z). \end{cases}$$

This implies that the univariate map $G(z) \to G(z)\Theta(z)$ assumes the form

$$(5.39\text{a})\qquad \left[\begin{array}{c|c} z & O \\ O & z^{-1} \end{array}\right] G_{j+1}(z) = G_j(z)\Theta_j(z),$$

where

$$(5.39\text{b})\qquad \Theta_j(z) = \left[\begin{array}{c|c} 1 & -k_j \\ k_j^* & 1 \end{array}\right] \left[\begin{array}{c|c} (1 + \frac{\tau_j^*}{2} z^{t_j} - \frac{\tau_j}{2} z^{-t_j}) & \frac{1}{2}(\tau_j^* z^{t_j} - \tau_j z^{-t_j}) \\ \frac{1}{2}(\tau_j^* z^{t_j} - \tau_j z^{-t_j}) & (1 - \frac{\tau_j^*}{2} z^{t_j} + \frac{\tau_j}{2} z^{-t_j}) \end{array}\right].$$

It is clear that the $\Theta(z)$ in (5.39b) may introduce negative powers of z in the intermediate $\gamma_j(z)$ and $\delta_j(z)$. However, the final result would not show any such irregularity. This irregularity can be explained nicely using a transmission line interpretation.

5.4. Transmission line interpretation and causality. The step j of Schur reduction on $\alpha_j(z)$ and $\beta_j(z)$ produces a two-port network section $\Theta_j^T(z)$, where z represents a delay element. Next, $\gamma_j(z)$ and $\delta_j(z)$ are applied to this network section, and that results in the outputs $\gamma_{j+1}(z)$ and $\delta_{j+1}(z)$. Since the network section in (5.39b) is noncausal, it results in a noncausal output. However, we can get a causal output by appropriately delaying it. Multiplying the resulting $\gamma_{j+1}(z)$ and $\delta_{j+1}(z)$ by z at each network section simply accomplishes this task.

Gohberg–Semencul-type representation and this approach. Gohberg and Semencul [GS72] created a striking formula for the inverse of a finite Toeplitz matrix. However, this formula can be used if and only if the second largest principal submatrix is also nonsingular, but Toeplitz matrices without this property are not uncommon. Gohberg and Krunipik resolved this issue by embedding the given Toeplitz matrix in a one-size larger nonsingular Toeplitz matrix and then deriving a solution from there. The solution is nonunique and requires working with a bigger matrix. Our procedure always finds a Gohberg–Semencul-type representation without starting with a bigger matrix.

Example. Next we will consider inverting a 3×3 nonsingular symmetric Toeplitz matrix whose 2×2 principal minor is zero. Consider the 3×3 symmetric Toeplitz matrix T with first row $[1 \quad 1 \quad 0]$. The starting generator

$$(5.40\text{a}) \qquad G_0(z) = \begin{bmatrix} \alpha_0(z) & | & \beta_0(z) \\ \gamma_0(z) & | & \delta_0(z) \end{bmatrix} = \begin{bmatrix} 1+z & | & z \\ 1 & | & 1 \end{bmatrix}.$$

Using (5.37a)-(5.37b), we get

$$k_0 = 0 \quad \text{and} \quad \Theta_0(z) = \begin{bmatrix} z & | & O \\ O & | & 1 \end{bmatrix},$$

which implies

$$(5.40\text{b}) \qquad G_1(z) = \begin{bmatrix} \alpha_1(z) & | & \beta_1(z) \\ \gamma_1(z) & | & \delta_1(z) \end{bmatrix} = \begin{bmatrix} z+z^2 & | & z \\ z & | & 1 \end{bmatrix}.$$

Then, using (5.39a)-(5.39b) we get

$$k_1 = 1, \quad t_1 = 1, \quad \text{and} \quad \Theta_1(z) = \begin{bmatrix} (1+z-z^{-1}) & | & (z-z^{-1}) \\ -(z-z^{-1}) & | & (1-z+z^{-1}) \end{bmatrix},$$

which implies

$$[\gamma_3(z) \quad | \quad \delta_3(z)] = [z^{-1} - 1 + z^2 \quad | \quad z^{-1} - z + z^2].$$

The causal generator for negative of the inverse of T must be $[z\gamma_3(z) \quad z\delta_3(z)] = [1 - z + z^3 \quad 1 - z^2 + z^3]$. A straightforward calculation verifies this claim. In this particular case it is not possible to find a Gohberg–Semencul formula, but our procedure provides us with one similar kind of formula.

6. Solution to linear equations. It is clear that the solutions to the Yule–Walker equations are obtained directly from the generators of the inverse. However, solution to a system of linear equations with an arbitrary right-hand side requires extending the notion BQT generating function (5.7) to

$$(5.41) \quad Q_b(z,w) = \frac{1}{1-zw^*}\begin{bmatrix} \alpha(z) & | & \beta(z) \\ \gamma(z) & | & \delta(z) \end{bmatrix}\begin{bmatrix} 1 & | & O \\ O & | & -1 \end{bmatrix}\begin{bmatrix} \alpha(w) & | & \beta(w) \\ \kappa(w) & | & \chi(w) \end{bmatrix}^*,$$

where $\alpha(z)$, $\beta(z)$, $\gamma(z)$, $\delta(z)$, $\kappa(z)$, and $\chi(z)$ are all polynomials (functions) in z.

The linear equation

$$Q\underline{x} = \underline{y}$$

for an arbitrary column vector $\underline{y}$ can be solved by choosing $\gamma(z) = \tilde{\gamma}_e(z) = \delta(z)$ (see (5.12c)) and $\kappa(z) = \tilde{\gamma}_e(z)y^{\#}(\underline{z}) = \chi(z)$, where $y(z) = [1 \quad z \quad z^2 \quad \cdots \quad z^n]\underline{y}$ and $y^{\#}(z) = z^n y^*(1/z^*)$. It turns out the Schur complements can be computed via trivial modification of the linear recursions in § 5.3. In fact, the $\Theta_j(z)$'s in (5.34)-(5.39b) remain the same, while a strongly regular step of Schur complementation is translated into two univariate maps $G^i(z) \to G^i(z)\Theta(z)$, $1 \leqq i \leqq 2$ of the form

$$(5.42) \qquad \begin{bmatrix} z & | & O \\ O & | & 1 \end{bmatrix} G^i_{j+1}(z) = G_j(z)\Theta_j(z).$$

Let us define the partitions

$$(5.43\text{a}) \qquad G^1_j(z) = \begin{bmatrix} \alpha_j(z) & | & \beta_j(z) \\ \gamma_j(z) & | & \delta_j(z) \end{bmatrix}$$

and

$$G_j^2(z) = \left[\begin{array}{c|c} \alpha_j(z) & \beta_j(z) \\ \kappa_j(z) & \chi_j(z) \end{array}\right]. \tag{5.43b}$$

In the singular case (see (5.38)-(5.39)) the univariate maps $G^i(z) \rightarrow G^i(z)\Theta(z)$ assume the form

$$\left[\begin{array}{c|c} z & O \\ O & z^{-1} \end{array}\right] G_{j+1}^i(z) = G_j^i(z)\Theta_j(z), \tag{5.44a}$$

where

$$\Theta_j(z) = \left[\begin{array}{c|c} 1 & -k_j \\ k_j^* & 1 \end{array}\right] \left[\begin{array}{c|c} (1 + \frac{\tau_j^*}{2} z^{t_j} - \frac{\tau_j}{2} z^{-t_j}) & \frac{1}{2}(\tau_j^* z^{t_j} - \tau_j z^{-t_j}) \\ \frac{1}{2}(\tau_j^* z^{t_j} - \tau_j z^{-t_j}) & (1 - \frac{\tau_j^*}{2} z^{t_j} + \frac{\tau_j}{2} z^{-t_j}) \end{array}\right]. \tag{5.44b}$$

7. Concluding remarks. A new approach has been developed that provides a unified framework for fast and completely recursive procedures for computing a modified triangular factorization of Hermitian Toeplitz and QT matrices. This recursive procedure is based on the fact that the Schur complement (or the block Schur complement) of the top left entry (or block) of a Toeplitz or a quasi-Toeplitz matrix is quasi-Toeplitz; except for Kailath and his associates (see [LK86], [BK88], and [Chu89]), other authors did not make use of this fact. This procedure does not require computation of inner products and is completely parallelizable. Determination of the size of a block factor is done by counting the number of repeated zeros at the origin of a polynomial. It also turns out that the block diagonal entries of D are even sized with an equal number of positive and negative eigenvalues.

Using the results on modified triangular factorization, we have extended the Schur complement-based approach of Chun [Chu89] for solutions to linear equations and inversion of strongly regular Toeplitz matrices to Hermitian Toeplitz and quasi-Toeplitz matrices with arbitrary rank profile. This leads to simultaneous derivation of Schur- and Levinson-type recursions. The inversion procedures work as long as the underlying matrices are nonsingular. Unlike all previous approaches (see Heinig and Rost [HR84] and Ben-Artzi and Shalom [BS86]), no special requirement on the rank profile is needed.

Since theoretical detection of a singularity and the number of consecutive zero minors require infinite precision, it is clear that a floating point implementation of the algorithm would contain serious problems. This raises the issue of devising a numerically sound general algorithm that would be able to deal with "nearly singular cases." Such a question goes beyond the scope of the paper and deserves a thorough investigation.

Acknowledgments. The authors thank Professor Hanoch Lev-Ari of Northeastern University for his constructive criticism and suggestions for improvement. The first author thanks Dr. Victor B. Lawrence, Dr. Joseph G. Kneuer, Dr. William H. Ninke, and Dr. Bryan D. Ackland of AT&T Bell Laboratories for their encouragement and support.

REFERENCES

[Bar69] E. J. Bareiss, *Numerical solution to linear equation with Toeplitz and vector Toeplitz matrices*, Numer. Math., 13 (1969), pp. 404–424.

[Ber68] E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

[BGY80] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun, *Fast solution ot Toeplitz systems of equations and computation of Padé approximants*, J. Algorithms, 1 (1980), pp. 259–295.

[BK88] Y. Bistritz and T. Kailath, *Inversion and factorization of non-Hermitian quasi-Toeplitz matrices*, Linear Algebra Appl., 98 (1988), pp. 77–121.

[BKP76] J. Bunch, L. Kauffman, and B. N. Parlett, *Decomposition of a symmetric matrix*, Numer. Math., 27 (1976), pp. 95–109.

[BS86] A. Ben-Artzi and T. Shalom, *On inversion of Toeplitz and close to Toeplitz matrices*. Linear Algebra Appl., 75 (1986), pp. 173–192.

[Bur75] J. P. Burg, *Maximum Entropy Spectral Analysis*, Ph.D. thesis, Dept. of Geophysics, Stanford University, Stanford, CA, 1975.

[Chu89] J. Chun, *Fast Array Algorithms for Structured Matrices*, Ph.D. thesis, Stanford University, Stanford, CA, June 1989.

[DGK85] P. Delsarte, Y. Genin, and Y. Kamp, *A Generalization of the Levinson Algorithm for Hermitian Toeplitz Matrices with Arbitrary Rank Profile*, IEEE Trans. Acoust. Speech Signal Process. 33 (1985), pp. 964–971.

[Dur60] J. Durbin, The fitting of time series models, Rev. Inst. Int. Stat., 28 (1960), pp. 233–243.

[DVK78] P. Dewilde, A. C. G. Vieira, and T. Kailath, *On a generalized Szegö–Levinson realization algorithm for optimal linear predictors based on a network synthesis approach*, IEEE Trans. Circuits Systems, 25 (1978), pp. 663–675.

[GB85] M. J. C. Gover and S. Barnett, *Inversion of Toeplitz matrices which are not strongly nonsingular*, IMA J. Numer. Anal., 5 (1985), pp. 101–110.

[Gra74] W. B. Gragg, *Matrix interpretations and applications of the continued fraction algorithm*, Rocky Mountain J. Math., 4 (1974), pp. 213–225.

[GS72] I. Gohberg and A. A. Semencul, *The inversion of finite Toeplitz matrices and their continual analogues*, Mat. Isled., 2 (1972), pp. 201–233.

[HR84] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-Like Matrices and Operators*, Akademie-Verlag, Berlin, 1984.

[Ioh82] I. S. Iohvidov, *Hankel and Toeplitz Matrices and Forms*, Birkhauser, Verlag, Basel, 1982.

[Kai87] T. Kailath, *Signal processing applications of some moment problems*, in Moments in Mathematics, H. Landau, ed., Proceedings of Symposia in Applied Mathematics, American Mathematical Society, Providence, RI, 1987, pp. 71–109.

[KBM86] T. Kailath, A. M. Bruckstein, and D. Morgan. *Fast matrix factorization via discrete transmission lines*, Linear Algebra Appl., 75 (1986), pp. 1–25.

[KVM78] T. Kailath, A. C. G. Vieira, and M. Morf, *Inverses of Toeplitz operators, innovations and orthogonal polynomials*, SIAM Rev. 20 (1978), pp. 106–119.

[LCC90] G. Labahn, D. K. Choi, and S. Cabay, *The inverses of block Hankel and block Toeplitz matrices*, SIAM J. Comput., 19 (1990), pp. 93–123.

[Lev47] N. Levinson, *The Wiener* RMS *error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[LK84] H. Lev-Ari and T. Kailath, *Lattice filter parameterization of non-stationary processes*, IEEE Trans. Inform. Theory, 30 (1984), pp. 2–16.

[LK86] ———, *Triangular factorization of structured Hermitian matrices*, Oper. Theory: Adv. Appl., 18 (1986), pp. 301–324.

[Mas69] J. L. Massey, *Shift-register synthesis and* BCH *decoding*, IEEE Trans. Inform. Theory, 15 (1969), pp. 122–127.

[Mor70] M. Morf, *Research Memorandum*, *Information Systems Laboratory*, Ph.D. thesis, Stanford University, Stanford, CA, 1970.

[Mor74] ———, *Fast Algorithms for Multivariable Systems*, Ph.D. thesis, Stanford University, Stanford, CA, August 1974.

[Pal90] D. Pal, *Fast Algorithms for Structured Matrices with Arbitrary Rank Profile*, Ph.D. thesis, Stanford University, Stanford, CA, May 1990.

[PLK88] S. Pombra, H. Lev-Ari, and T. Kailath, *Levinson and Schur algorithms for Toeplitz matrices with singular minors*, in Proc. ICASSP-88, New York, (1988), pp. 1643–1646.

[RG77] J. Le Roux and C. Gueguen. *A fixed point computation of partial correlation coefficients*, IEEE Trans. Acoust. Speech Signal. Process., 25 (1977), pp. 257–259.

[Ris73] J. Rissanen, *Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with applications to factoring positive polynomials*, Math. Comput., 27 (1973), pp. 147–154.

[Sch17] I. Schur, *Uber Potenzreihen, die in Innern des Einheitskreises Beschrankt Sind*, J. Reine Angew. Math., 147 (1917), pp. 205–232.

[Sug86] Y. Sugiyama, *An algorithm for solving discrete time Wiener-Hopf equations based on Euclid's algorithm*, IEEE Trans. Inform. Theory, 32 (1986), pp. 394–409.

[Tre64] W. F. Trench, *An algorithm for inversion of finite Toeplitz matrices*, SIAM J. Appl. Math., 12 (1964), pp. 515–522.

[Whi63] P. Whittle, *Prediction and Regulation by Linear Least-Squares Methods*. Van Nostrand, Princeton, NJ, 1963.

[Zoh69] S. Zohar, *The algorithm of W. F. Trench*, J. Assoc. Comput. Mach., 16 (1969), pp. 592–601.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1043–1047, October 1993

008

LENGTH BOUNDS FOR SINGULAR VALUES OF SPARSE MATRICES*

CHARLES R. JOHNSON† AND PETER M. NYLEN‡

Abstract. For an m-by-n matrix A with singular values $\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_n$, the following is observed. Concatenate all rows of A having nonzero entries in the ith column and calculate the Euclidean length ℓ_i of the result. Let $\hat{\ell}_1 \geqq \hat{\ell}_2 \geqq \cdots \geqq \hat{\ell}_n$ be a rearrangement of the numbers ℓ_i, $i = 1, \ldots, n$. Then $\sigma_i \leqq \hat{\ell}_i$, $i = 1, \ldots, n$. Examples show that $\hat{\ell}_1$ can be smaller than the bound for σ_1 obtained by applying Gersgorin's theorem to A^*A.

Key words. singular values, sparse matrix

AMS subject classifications. 15A18, 15A45, 15A60, 65F35, 65F50

Let $A = (a_{ij})$ be an m-by-n complex matrix with singular values $\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_n \geqq 0$. It is well known that σ_1 is at least the greatest row length or column length of A, and that σ_1 is no more than the geometric mean of the largest absolute row and column sums of A or the square root of the largest absolute row sum of A^*A [HJ]. Our interest here is in describing selections of entries, from A, whose Euclidean lengths give upper bounds for the singular values of A. These will be of most interest when A is relatively sparse. In this event, our bounds can be better than standard known bounds. Of course, if the columns of A are combinatorially orthogonal, then the column lengths will be the singular values, and, if A is dense, then the length of A, thought of as a vector (the Frobenius norm), will be an upper bound for σ_1. Both of these are special cases of our observations.

Our most succinct result is the following. For $i = 1, \ldots, n$, define ℓ_i as the Euclidean length of the concatenation of all rows of A having nonzero entries in the ith column. (Equivalently, ℓ_i is the Frobenius norm of the submatrix lying in the rows of A in which column i has nonzero entries.) Reorder the ℓ_i's in descending order to obtain the sequence $\hat{\ell}_1 \geqq \hat{\ell}_2 \geqq \cdots \geqq \hat{\ell}_n \geqq 0$. For example, if

$$A = \begin{bmatrix} a_{11} & 0 & a_{13} & 0 \\ a_{21} & a_{22} & 0 & 0 \\ 0 & a_{32} & 0 & 0 \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix},$$

then

$$\ell_1 = (|a_{11}|^2 + 0^2 + |a_{13}|^2 + 0^2 + |a_{21}|^2 + |a_{22}|^2 + 0^2 + 0^2)^{1/2},$$

$$\ell_2 = (|a_{21}|^2 + |a_{22}|^2 + 0^2 + 0^2 + 0^2 + |a_{32}|^2 + 0^2 + 0^2)^{1/2},$$

$$\ell_3 = (|a_{11}|^2 + |a_{13}|^2 + |a_{43}|^2 + |a_{44}|^2)^{1/2},$$

$$\ell_4 = (|a_{43}|^2 + |a_{44}|^2)^{1/2},$$

* Received by the editors March 4, 1991; accepted for publication (in revised form) January 21, 1992. This work was supported in part by National Science Foundation grant DMS 90-00839, Office of Naval Research contract N00014-90-J-1739, and a North Atlantic Treaty Organization travel grant.

† Department of Mathematics, The College of William and Mary, Williamsburg, Virginia 23187-8795 (crjohn@wmvm1.bitnet).

‡ Department of Mathematics, Auburn University, Auburn, Alabama 36849 (pmnylen@auducvax.bitnet).

and

$$\hat{\ell}_1 = \max\{\ell_1, \ell_2, \ell_3, \ell_4\}; \hat{\ell}_2 = \max\{\ell_1, \ell_2, \ell_3, \ell_4\}/\{\hat{\ell}_1\},$$
$$\hat{\ell}_3 = \min\{\ell_1, \ell_2, \ell_3, \ell_4\}/\{\hat{\ell}_4\}; \hat{\ell}_4 = \min\{\ell_1, \ell_2, \ell_3, \ell_4\}.$$

The result then is the following theorem.

THEOREM. *For any m-by-n complex matrix A, we have*

$$\sigma_i(A) \leqq \hat{\ell}_i(A), \quad i = 1, \dots, n.$$

Of course the theorem may be applied to A^*, whose nonzero singular values are the same as those of A, to obtain also

$$\sigma_i(A) \leqq \hat{\ell}_i(A^*), \quad i = 1, \dots, m,$$

so that we have

$$\sigma_i(A) \leqq \min\{\hat{\ell}_i(A), \hat{\ell}_i(A^*)\}, \qquad i = 1, \dots, \min\{m, n\}.$$

Although the theorem produces rather good bounds, especially for certain sparsity patterns, perhaps more interesting is the technique of proof. We adjoin rows to A to produce a matrix A' with orthogonal columns and notice that the singular values of a matrix with orthogonal columns are its column lengths. As deletion of rows from a matrix cannot increase any singular value because of interlacing [HJ], the column lengths of A', properly ordered, bound the singular values of A. The simplest construction of such a matrix A' is as follows. First adjoin rows to A with two nonzero entries to make the first column orthogonal to all subsequent columns. For each pair of nonzero entries of the form $(a_{k1}, a_{kj}), j \neq 1$, adjoin the row to A with first entry $-\bar{a}_{kj}$, jth entry $\bar{a}_{k1}$, and all other entries zero. Note that the length of the new first column is $\ell_1(A)$, as it contains all its original nonzero entries plus one copy of (the negative conjugate of) each nonzero entry "hit" by a nonzero entry of column one of A in the inner product of that column with subsequent columns. Now, "fix up" the orthogonality of the second column with all subsequent columns in exactly the same way and note that the orthogonality between column one and subsequent columns is not disturbed. Now, the length of column two is $\ell_2(A)$. Continue in the same manner, adjusting column three versus the following, and so on, producing A'. For instance, if

$$A = \begin{bmatrix} a_{11} & 0 & a_{13} & 0 \\ a_{21} & a_{22} & 0 & 0 \\ 0 & a_{32} & 0 & 0 \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix}, \quad \text{then} \quad A' = \begin{bmatrix} a_{11} & 0 & a_{13} & 0 \\ a_{21} & a_{22} & 0 & 0 \\ 0 & a_{32} & 0 & 0 \\ 0 & 0 & a_{43} & a_{44} \\ -\bar{a}_{22} & \bar{a}_{21} & 0 & 0 \\ -\bar{a}_{13} & 0 & \bar{a}_{11} & 0 \\ 0 & 0 & -\bar{a}_{44} & \bar{a}_{43} \end{bmatrix}.$$

This completes a proof of the theorem, but we note that the same technique may be used to exploit special structure, such as sparsity or relations among entries. For example, if

$$A = \begin{bmatrix} a & 0 & 0 & b \\ 0 & a & 0 & c \\ 0 & 0 & a & d \end{bmatrix},$$

then for

$$A' = \begin{bmatrix} a & 0 & 0 & b \\ 0 & a & 0 & c \\ 0 & 0 & a & d \\ -\bar{b} & -\bar{c} & -\bar{d} & \bar{a} \\ \bar{c} & -\bar{b} & 0 & 0 \\ 0 & \bar{d} & -\bar{c} & 0 \\ \bar{d} & 0 & -\bar{b} & 0 \end{bmatrix},$$

we have $\sigma_1(A) \leqq (|a|^2 + |b|^2 + |c|^2 + |d|^2)^{1/2}$. In fact, equality holds.

We note that the theorem can provide better bounds than conventional estimates. For example, if

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 2 \\ 1 & 0 & 0 & 2 \end{bmatrix},$$

then the maximum absolute row and column sums of A are three and four, so that

$$\sigma_1(A) \leqq \sqrt{12}.$$

The maximum absolute row sums of A^*A and AA^* are 12 and 11, so that

$$\sigma_1(A) \leqq \sqrt{11}.$$

However, $\ell_1(A) = \ell_2(A) = \ell_3(A) = \ell_4(A) = \sqrt{4 + 4 + 1 + 1} = \sqrt{10}$ so that, by the theorem,

$$\sigma_1(A) \leqq \sqrt{10}.$$

In fact, $\sigma_1(A) = \sqrt{10}$; for

$$x = \begin{pmatrix} 2 \\ 1 \\ 2 \\ 4 \end{pmatrix},$$

the length of Ax is $5\sqrt{10}$, while the length of x is 5.

This example suggests one of the patterns for which the theorem provides nice bounds. We call an n-by-n matrix $A = (a_{ij})$ *k-cyclic* if all its nonzero entries are contained among $a_{11}, a_{22}, \dots, a_{nn}, a_{12}, a_{23}, \dots, a_{n1}, \dots, a_{1k}, a_{2,k+1}, \dots, a_{n,k-1}$. The last example is two-cyclic. If A is k-cyclic, then the nonzero entries in column one of A are in positions $1, n - k + 2, \dots, n$, in column two are in positions $1, 2, n - k + 3, \dots, n, \dots$, and in column n are in positions $n - k + 1, n - k + 2, \dots, n$. The largest singular value of a k-cyclic matrix is then no more than the maximum of the lengths of the concatenations of each of these n sets of k rows.

The computational complexity of obtaining the numbers $\hat{\ell}_1 \geqq \cdots \geqq \hat{\ell}_n$ compares favorably with any bound using A^*A or AA^*, such as

$$g(A) \equiv (\text{maximum absolute row sum of } AA^*)^{1/2},$$

as $\ell_1^2, \dots, \ell_n^2$ may be found at less expense than AA^* itself. In fact, finding $\ell_1^2, \dots, \ell_n^2$ requires little more effort than finding the diagonal of AA^*. For diag (AA^*) =

$(b_1, \dots, b_n)$, we have

$$\ell_j^2 = \sum_{i:a_{ij}\neq 0} b_i .$$

For real matrices, the $\hat{\ell}_i$ are slightly more costly to compute than the upper bound for σ_1 given by

$$\begin{aligned} rc(A) = (&\text{maximum absolute column sum of } A)^{1/2} \\ (&\text{maximum absolute row sum of } A)^{1/2}. \end{aligned}$$

For complex matrices, the $\hat{\ell}_i$ have the advantage. In any event, neither $g(A)$ nor $rc(A)$ gives smaller bounds for the other singular values, as the $\hat{\ell}_i$ do.

It is easy to see that $rc(A) \geqq g(A)$ for all A and that

$$n^{-1/2} g(A) \leqq \hat{\ell}_1(A) \leqq n^{1/2} g(A)$$

and

$$n^{-1/2} rc(A) \leqq \hat{\ell}_1(A) \leqq n^{1/2} rc(A).$$

We may verify that the upper bound on $\hat{\ell}_1$ cannot be improved by considering the matrix

$$\begin{bmatrix} 1 & 1 & \cdots & 1 & 0 \\ & & & & 1 \\ & & 0 & & 1 \\ & & & & \vdots \\ & & & & 1 \end{bmatrix}.$$

The matrix

$$I + \varepsilon J,$$

with $\varepsilon > 0$ small and J the matrix of all ones, shows that the lower bound cannot be improved.

We have not characterized those matrices for which $\hat{\ell}_1(A) \leqq g(A)$ (or $\hat{\ell}_1(A) \leqq rc(A)$), although there are many. We conclude with the results of a numerical experiment comparing $\hat{\ell}_1$ and g. The experiment was performed (for each size and nonzero fraction) by generating 100 random zero-nonzero patterns, generating 25 random matrices (with entries chosen from $\{-4, -3, \dots, 5\}$ in each zero-nonzero class), and evaluating the number of times $\hat{\ell}_1(\cdot)$ (respectively, $g(\cdot)$) is minimal, and ties are counted in both columns. The "flops" column gives a measure, in FORTRAN flops, of total calculation costs. The "bad patterns" column records the number of zero-nonzero patterns for which

TABLE 1
10×10.

Fraction nonzero	$\hat{l}_1$ wins	Flops	Bad patterns	g wins	Flops	rc flops
.1	2103	498207	0	575	5225000	452500
.15	1923	509253	0	636	5225000	452500
.2	1649	519935	1	849	5225000	452500
.25	1448	530644	1	1093	5225000	452500
.3	1056	542443	4	1469	5225000	452500

TABLE 2
20×20.

Fraction nonzero	$\hat{l}_1$ wins	Flops	Bad patterns	g wins	Flops	rc flops
.05	2152	1994005	0	451	40950000	1902500
.1	1894	2040804	0	636	40950000	1902500
.15	1377	2085634	0	1150	40950000	1902500
.2	761	2131355	3	1763	40950000	1902500
.25	304	2172219	17	2205	40950000	1902500
.3	40	2223201	77	2462	40950000	1902500

$g(\cdot) \leqq \hat{\ell}_1(\cdot)$ for all 25 matrices. As can be seen in Tables 1 and 2, $\hat{\ell}_1$ is generally better (and much cheaper) for a high degree of sparsity and is not entirely dominated even for lower degrees of sparsity.

REFERENCE

[HJ] R. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1048–1060, October 1993

SOLUTION OF P_0-MATRIX LINEAR COMPLEMENTARITY PROBLEMS USING A POTENTIAL REDUCTION ALGORITHM*

PANOS M. PARDALOS[†], YINYU YE[‡], CHI-GEUN HAN[§], AND JOHN A. KALISKI[¶]

Abstract. The convergence result [Y. Ye and P. M. Pardalos, *Linear Algebra Appl.*, 152 (1991), pp. 3–17] of a potential reduction algorithm for the P-matrix linear complementarity problem (LCP) is extended to the P_0-matrix LCP. Computational experience with PSD-matrix, P-matrix, and P_0-matrix LCPs is presented to reinforce the theoretical development. These test problems include random test problems, the worst case examples of Murty and Fathi, and various LCPs arising in engineering problems. How row and column scaling may improve the theoretical and practical efficiency of the algorithm is also illustrated.

Key words. linear complementarity problem, P_0-matrix, interior point algorithms, test problems

AMS subject classification. 90C33

1. Introduction.

For a given n by n rational matrix M and a given n rational vector q, the linear complementarity problem, $LCP(M, q)$, is that of finding an (x, y) such that

$$y = Mx + q, \quad x \geq 0, \quad y \geq 0, \quad x^T y = 0,$$

or proving that no such (x, y) exists. Economic equilibrium problems, game theory problems, numerical solutions of differential equations, and the necessary and sufficient conditions of optimality in optimization are all sources of linear complementarity problems ([2], [3], [4], [16]).

Some of the earliest proposed algorithms for solving LCPs include the principal pivoting method of Cottle and Dantzig, Lemke's complementary pivot algorithm, Mangasarian's linear programming algorithm, Murty's Bard-type method, and Van der Heyden's variable dimension algorithm (e.g., [2], [3], [12], [16]). The recent book by Murty [16] gives a complete presentation of these algorithms. Most of these algorithms are guaranteed to solve only LCPs with some special structure imposed on the matrix M. When the matrix M has no special structure, the LCP becomes computationally very difficult with the only existing solution techniques being of the enumerative type or of an equivalent integer programming formulation ([17], [18]).

Although the general LCP is NP-hard, some classes of LCPs can be solved by polynomial time algorithms. For example, when M is positive semi-definite (PSD), the LCP is a convex quadratic program and can be solved by the ellipsoid algorithm or several interior point algorithms (e.g., see Kojima et al. [10]). Other classes are discussed in [12] and [24].

An LCP is called a P-matrix (P_0-matrix) LCP when the associated matrix M is a P-matrix (P_0-matrix). An n by n matrix M is a P-matrix (P_0-matrix) if and only

* Received by the editors July 19, 1990; accepted for publication (in revised form) December 26, 1991.

[†] Department of Industrial and Systems Engineering, 303 Weil Hall, University of Florida, Gainesville, Florida 32611 (pardalos@math.ufl.edu).

[‡] Department of Management Sciences, University of Iowa, Iowa City, Iowa 52242. The research of this author was supported in part by National Science Foundation grant DDM-8922636 and the Iowa Business School Summer grant (BBUYINVA@UIAMVS.BITNET).

[§] Department of Computer Science, Pennsylvania State University, University Park, Pennsylvania 16802 (cghan@nms.Kyunghee.ac.Kr).

[¶] Department of Management Sciences, University of Iowa, Iowa City, Iowa 52242.

if all its principal minors are positive (nonnegative) (e.g., see Berman and Plemmons [1], Fiedler [6], and Fiedler and Pták [7]). It is well known that the $LCP(M,q)$ has a unique solution for every q if and only if M is a P-matrix. At present, no polynomial time algorithm is known either to test whether a general square matrix is a P-matrix, or to solve a P-matrix LCP.

Some attempts to understand the complexity of a P-matrix LCP include the results by Megiddo [14] and by Solow, Stone, and Tovey [19]. Megiddo proved that if it is NP-hard to solve an LCP with a P-matrix, then $NP = coNP$. Another result is that the P_0-matrix LCP is NP-complete [10].

In this paper we extend our convergence result in [24] the potential reduction algorithm for the row-sufficient-matrix (a special P_0-matrix) LCP to the general P_0-matrix LCP. We also present computational experience with P-matrix and P_0-matrix LCPs. The paper is organized as follows. First the potential reduction algorithm is reviewed and analyzed. In particular, we prove that a P_0-matrix LCP has a positive "condition number" as does the row-sufficient-matrix LCP. The condition number serves as a measure of the convergence speed for the potential reduction algorithm for solving the LCP. After this analysis, we present some computational results with a large class of test problems. These test problems include the worst case examples by Murty [15] and Fathi [5], and some P-matrix and P_0-matrix LCPs originating from engineering problems. Finally, we illustrate how row and column scaling can be used to improve the theoretical and practical performance of the algorithm.

Given the $LCP(M,q)$, we assume throughout the remainder of this paper that the set $\Omega^+ = \{(x,y) : x > 0,\ y = Mx + q > 0\}$ is nonempty. An LCP with empty Ω^+ can be transferred to an equivalent or related LCP with nonempty Ω^+, see Kojima et al. [10], [11]. We also denote the associated feasible domain by $\Omega = \{(x,y) : x \geq 0,\ y = Mx + q \geq 0\}$.

2. A potential reduction algorithm.

We now review the potential function and potential reduction algorithm for the LCP. The potential function

$$\phi(x,y) = \rho \log(x^T y) - \sum_{j=1}^{n} \log(x_j y_j) \tag{2.1}$$

$$= (\rho - n)\log(x^T y) - \sum_{j=1}^{n} \frac{\log(x_j y_j)}{\log(x^T y)},$$

where $\rho > n$, is used with an interior feasible solution (x,y). This function first appeared in Todd and Ye [20]. The gradients of the potential function are

$$\nabla_x \phi(x,y) = \frac{\rho}{x^T y} y - X^{-1}e \quad \text{and} \quad \nabla_y \phi(x,y) = \frac{\rho}{x^T y} x - Y^{-1}e,$$

where e is the vector of all ones. Starting from an interior point (x^0, y^0) which satisfies

$$\phi(x^0, y^0) \leq O(\rho L),$$

where L is the size of the input data M and q, the potential reduction algorithm generates a sequence of interior feasible solutions $\{x^k, y^k\}$ terminating at a point such that

$$\phi(x^k, y^k) \leq -(\rho - n)L.$$

It can be verified at this point that

$$(x^k)^T y^k \leq 2^{-L},$$

and an exact solution to the LCP can be obtained in $O(n^3)$ additional operations.

To achieve a potential reduction, we used the scaled gradient projection method (see Kojima, Megiddo, and Ye [11] and Ye [23]). We solve the following linear program subject to an ellipsoid constraint at the kth iteration:

$$\begin{aligned} \text{minimize} \quad & \nabla_x \phi(x^k, y^k)\delta x + \nabla_y \phi(x^k, y^k)\delta y \\ \text{subject to} \quad & \delta y = M\delta x, \\ & \|(X^k)^{-1}\delta x\|^2 + \|(Y^k)^{-1}\delta y\|^2 \leq \beta^2 < 1. \end{aligned}$$

The scaled gradient projection vector to the null space of the scaled equality constraints is

$$p^k = \begin{pmatrix} p_x^k \\ p_y^k \end{pmatrix} = \begin{pmatrix} \frac{\rho}{\Delta^k} X^k(y^k + M^T\pi^k) - e \\ \frac{\rho}{\Delta^k} Y^k(x^k - \pi^k) - e \end{pmatrix}, \tag{2.2}$$

where

$$\pi^k = ((Y^k)^2 + M(X^k)^2M^T)^{-1}(Y^k - MX^k)\left(X^k y^k - \frac{\Delta^k}{\rho} e\right), \tag{2.3}$$

and $\Delta^k = (x^k)^T y^k$, X^k (Y^k) denotes the diagonal matrix of x^k (y^k). Then, for some $\beta > 0$ we assign

$$x^{k+1} = x^k - \beta X^k p_x^k / \parallel p^k \parallel \quad \text{and} \quad y^{k+1} = y^k - \beta Y^k p_y^k / \parallel p^k \parallel .$$

By choosing

$$\beta = \min\left(\frac{\|p^k\|}{\rho + 2}, \frac{1}{2}\right) \leq \frac{1}{2}, \tag{2.4}$$

we can prove

$$\phi(x^{k+1}, y^{k+1}) - \phi(x^k, y^k) \leq -\alpha(\|p^k\|^2), \tag{2.5}$$

where $\alpha(\|p^k\|^2) = \|p^k\|^2/(2(\rho + 2))$ if $\|p^k\|^2 \leq (\rho + 2)^2/4$, and $\alpha(\|p^k\|^2) = (\rho + 2)/8$ otherwise.

The algorithm can be simply stated.

Potential Reduction Algorithm

```
Given $x^0, y^0 > 0$ and $y^0 = Mx^0 + q$;
while $(x^k)^T y^k \geq 2^{-L}$ do
  begin
    compute $\pi^k$ of (2.3) and $p^k$ of (2.2), and select $\beta$;
    let $x^{k+1} = x^k - \beta X^k p_x^k / \parallel p^k \parallel$ and $y^{k+1} = y^k - \beta Y^k p_y^k / \parallel p^k \parallel$;
    $k = k + 1$;
  end
```

From (2.5) we see that $\|p^k\|^2$ partially determines the potential reduction at the kth iteration of the algorithm. Note that the greater the value of $\|p^k\|^2$, the larger the potential reduction. Let

$$g(x,y) = \frac{\rho}{\Delta}Xy - e$$

and

$$H(x,y) = 2I - (XM^T - Y)(Y^2 + MX^2M^T)^{-1}(MX - Y).$$

Then

$$\|p^k\|^2 = g^T(x^k, y^k)H(x^k, y^k)g(x^k, y^k).$$

We then use $\|g(x,y)\|_H^2$ to denote $g^T(x,y)H(x,y)g(x,y)$.

Ye and Pardalos [24] defined the condition number for the $LCP(M,q)$ as

$$\gamma(M,q) = \inf\{\|g(x,y)\|_H^2 : x^Ty \geq 2^{-L},\ \phi(x,y) \leq O(\rho L) \text{ and } (x,y) \in \Omega^+\}.$$

There is a sequence of propositions for $\gamma(M,q)$.

PROPOSITION 1 (Kojima, Megiddo, and Ye [11]). *Let $\rho \geq 2n + \sqrt{2n}$. Then, for M being a* PSD *matrix and any $q \in R^n$,*

$$\gamma(M,q) \geq 1.$$

PROPOSITION 2 (Ye [22]). *Let $\rho \geq 3n + \sqrt{2n}$. Then, for M being a P-matrix and any $q \in R^n$,*

$$\gamma(M,q) \geq \min(n\theta/|\lambda|, 1),$$

where λ is the least eigenvalue of $(M + M^T)/2$, and θ is the P-matrix number of M^T, i.e.,

$$\theta = \min_{x \neq 0}\left\{\max_j \frac{x_j(M^Tx)_j}{\|x\|^2}\right\}.$$

PROPOSITION 3 (Ye and Pardalos [24]). *Let $\rho > n$ and be fixed. Then, for M being a row-sufficient matrix and $\{(x,y) \in \Omega^+ : \phi(x,y) \leq O(\rho L)\}$ being bounded,*

$$\gamma(M,q) > 0.$$

In this paper, we prove the following proposition.

PROPOSITION 4. *Let $\rho > n$ and be fixed. Then, for M being a P_0-matrix and $\{(x,y) \in \Omega^+ : \phi(x,y) \leq O(\rho L)\}$ being bounded,*

$$\gamma(M,q) > 0.$$

Proof. In Proposition 4 of [24] it was shown that $\{(x,y) \in \Omega^+ : x^Ty \geq 2^{-L},\ \phi(x,y) \leq O(\rho L)\}$ is contained in a closed and bounded set $\{(x,y) \in \Omega : x \geq \bar{\epsilon}e,\ y \geq \bar{\epsilon}e,\ \phi(x,y) \leq O(\rho L)\}$ for some fixed number $\bar{\epsilon} > 0$. Using this compactness, we need only to show that for any $(x,y) \in \Omega^+$

$$\|g(x,y)\|_H^2 > 0.$$

The proof is by contradiction. Suppose

$$\|g(x,y)\|_H = \left\| \begin{pmatrix} \frac{\rho}{\Delta} X(y + M^T\pi) - e \\ \frac{\rho}{\Delta} Y(x - \pi) - e \end{pmatrix} \right\| = 0,$$

then we must have

$$XM^T\pi + Y\pi = 0. \tag{2.6}$$

Let $\pi \neq 0$ since otherwise clearly $\|g(x,y)\|_H > 0$. However, M^T being a P_0-matrix implies that there exists an index j such that

$$\pi_j \neq 0 \quad \text{and} \quad \pi_j(M^T\pi)_j \geq 0.$$

This leads to

$$y_j(\pi_j)^2 + x_j\pi_j(M^T\pi)_j > 0,$$

which is a contradiction to equality (2.6). □

Note that $\phi(x,y) \leq (\rho - n)L$ implies that $x^Ty \leq \exp(L)$. Hence,

$$\{(x,y) \in \Omega^+ : \phi(x,y) \leq (\rho - n)L\} \subset \{(x,y) \in \Omega : x^Ty \leq \exp(L)\}, \tag{2.7}$$

and the boundedness of $\{(x,y) \in \Omega : x^Ty \leq \exp(L)\}$ guarantees the boundedness of the potential level set. It has been shown that the primal-dual algorithm of Kojima et al. [10] solves the P_0-matrix LCP if $\{(x,y) \in \Omega : x^Ty \leq \exp(L)\}$ is bounded. Here, from Theorem 1 of Ye and Pardalos [24], Proposition 4 indicates that the potential reduction algorithm solves the P_0-matrix LCP if the potential level set is bounded. Note that in general the boundedness of x^Ty does not imply the boundedness of $\phi(x,y)$. For example, let $x_1y_1 = 1$ and x_2y_2 approach zero. Then, $x_1y_1 + x_2y_2$ is bounded from above, but $\rho\log(x_1y_1 + x_2y_2) - \log(x_1y_1) - \log(x_2y_2)$ approaches infinity. However, we believe that these two sets are closely related.

3. Computational results. The potential reduction algorithm was implemented and tested on an IBM 3090-600S computer with Vector Facilities using the VS Fortran compiler. We used double precision for all computations to solve a variety of test problems with M being PSD, a P-matrix, and a P_0-matrix. LCPs with PSD matrices include the examples by Murty, Fathi, and random PSD matrices.

In general, an initial feasible solution may not be readily available for the potential reduction algorithm. However, many techniques have been proposed to overcome this difficulty (i.e., [10], [11]). For example, if M is a P_0-matrix, then we can let

$$M' = \begin{pmatrix} M & e - Me - q \\ 0 & 1 \end{pmatrix}, q' = (q^T, 0)^T.$$

For $LCP(M', q')$, $x^0 = e$ and $y^0 = e$ is an interior feasible solution. Also M' remains a P_0-matrix. If a solution exists for $LCP(M, q)$, then it is also a solution for $LCP(M', q')$, and vice versa.

The subroutines DPPF and DPPS of ESSL were used for solving the linear system (2.3). The DPPF factorizes the symmetric positive definite system using Gaussian elimination and the DPPS solves the factorized system. $x^Ty < \epsilon = 10^{-5}$ was used as a stopping criterion and ρ is set to $n^{1.5}$. The step size is chosen as 99% of the way to the boundary (see also [8]). All averages were obtained from five problems in each case and the CPU times are given in seconds. All entries of matrices and vectors used in the tests were generated randomly using the random number generator subroutine DURAND of the ESSL (if their elements are not explicitly described). Note that the notation $c_1 \leq v \leq c_2$ for a vector v means that value of v_i for $i = 1, \ldots, n$ is chosen randomly between c_1 and c_2.

3.1. Positive semi-definite matrices. First, we consider LCPs with M being a positive semi-definite matrix. We start with the two classical examples by Murty and Fathi and then solve random convex LCPs.

3.1.1. Murty's matrix. Murty [15] has shown that the computational requirements of two complementary pivot type algorithms exhibit an exponential growth in terms of the dimension of the LCP in the worst case. Murty's example has data:

$$M = \begin{pmatrix} 1 & & & & \\ 2 & 1 & & 0 & \\ 2 & 2 & 1 & & \\ . & . & . & . & \\ 2 & . & . & 2 & 1 \end{pmatrix}, q = (-1, \ldots, -1)^T.$$

The starting point is

$$x^0 = (2, 1, \ldots, 1), y^0 = (1, 4, 6, 8, \ldots, 2n),$$

and a solution of $LCP(M, q)$ is given by

$$x^* = (1, 0, \ldots, 0), y^* = (0, 1, \ldots, 1).$$

Note that the matrix M is triangular positive definite and that can be solved very efficiently by a backward substitution method.

3.1.2. Fathi's matrix. Fathi's example [5] has data:

$$M = \begin{pmatrix} 1 & 2 & 2 & . & 2 \\ 2 & 5 & 6 & . & 6 \\ 2 & 6 & 9 & . & 10 \\ 2 & 6 & 10 & . & . \\ . & . & . & . & . \\ 2 & 6 & 10 & . & 4(n-1)+1 \end{pmatrix}, q = (-1, \ldots, -1)^T.$$

The starting point is

$$x^0 = (1, 1, \ldots, 1), y^0 = Mx^0 + q,$$

and a solution of $LCP(M, q)$ is given by

$$x^* = (1, 0, \ldots, 0), y^* = (0, 1, \ldots, 1).$$

Note that the matrix M is PSD. The following Table 1 summarizes the number of iterations and CPU times required to solve the two examples for different values of n.

TABLE 1
Murty's and Fathi's matrices.

	Murty		Fathi	
n	Itr.	CPU	Itr.	CPU
100	14	0.35	16	0.57
200	15	2.19	18	3.96
300	15	7.62	18	12.88
400	15	16.14	17	28.56
500	15	29.92	18	56.02

3.1.3. Random positive semi-definite matrix. Next, we generated LCPs with M being a random positive semi-definite matrix. The matrix M has the form

$$M = U^T \Lambda U,$$

where $U \in R^{n\times n}$ is a Householder matrix of a random vector $r, 0 \le r_i \le 1, i = 1, \dots, n$ and $\Lambda = \text{diag}(\lambda_i), 0 \le \lambda_i \le 11, i = 1, \dots, n$. We choose a complementary solution (x^*, y^*) that satisfies: $y_i^* \ge 0$, $(y^* + Me)_i > 0$, $0 \le x_i^* \le 10, i = 1, \dots, n$; ${x^*}^T y^* = 0$, and $q = y^* - Mx^*$. Note that since M is a PSD matrix, $e^T Me \ge 0$. Hence, we have at least one positive element of Me (theoretically all entries of Me can be zero, but we do not expect that situation in practice), i.e., we have at least one index i such that $y_i^* = 0$. The initial solutions are chosen, respectively, as $x^* + e$ and $y^* + Me$. From the computational results (Table 2), we see that the number of iterations is quite insensitive to the size of the problem. Also note that the number of iterations is approximately the same among different problems with the same dimension. This behavior may be expected from our theoretical findings. The bound for the "condition number" of PSD LCPs is independent of both the data and the problem's size. The convergence behavior is also similar to our previous results with solving convex quadratic programs [8].

TABLE 2
Random positive semi-definite matrix.

n	Avg. Itr.	Avg. CPU
100	16.0	0.57
200	16.0	3.51
300	16.0	11.51
400	17.0	28.32
500	17.0	52.75

3.2. P-matrices. In this section, we solve LCPs with a P-matrix, where the matrix M is given by

$$M = \begin{pmatrix} P_1 & \alpha Q \\ 0 & P_2 \end{pmatrix},$$

where $P_1 \in R^{n_1\times n_1}$ and $P_2 \in R^{n_2\times n_2}$ are positive definite matrices, α is a positive scalar and $Q \in R^{n_1\times n_2}$. In our test problems, $n_1 = 2n/5$ and $n_2 = 3n/5$ and $P_k = A_k^T A_k$, where $A_k = \{a_{ij}^k\} \in R^{n_k\times n_k}$ and $0 < a_{ij}^k < 1$ for $k = 1, 2$. The matrix Q was given by

$$\begin{aligned} 0 < q_{ij} < 1 \quad & \text{if } i \neq j, \\ 1 < q_{ij} < 12 \quad & \text{if } i = j. \end{aligned}$$

Note that for α sufficiently large, the number of negative eigenvalues of $M + M^T$ is approximately equal to $\min(n_1, n_2)$. This may be seen as follows. Consider the following matrix M'

$$M' = \begin{pmatrix} 0 & \alpha I & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

$M' + (M')^T$ will have exactly n (the dimension of I) negative eigenvalues. As $\alpha \to \infty$, the matrix M will in block form approach the form of M' with a similar transformation.

Therefore, the parameter α controls, in a sense, the nonconvexity of the LCP. For the P-matrix LCP, it controls the condition number discussed previously. This control may be clearly seen in a simple example given by Mathias [13]. Consider

$$M = \begin{pmatrix} 1 & 2\alpha \\ 0 & 1 \end{pmatrix},$$

where $\alpha > 1$. The P-matrix number and the least eigenvalue of $\frac{1}{2}(M + M^T)$ are, respectively,

$$\theta = \frac{1}{2}\left(1 - \frac{\alpha}{\sqrt{\alpha^2+1}}\right) \quad \text{and} \quad \lambda = 1 - \alpha.$$

By Proposition 2, the condition number of the LCP is

$$\gamma(M, q) \geq \min(n\theta/|\lambda|, 1) = O(1/\alpha^3).$$

We see that as α increases, the condition number deteriorates.

Matrix scaling is an important technique that is commonly used in optimization to improve the practical efficiency and numerical behavior of algorithms [21]. Frequently a matrix M is scaled as follows:

$$\bar{M} = EMD,$$

where E and D are diagonal matrices with positive diagonal entries. Given this scaling and an $LCP(M, q)$, we may solve an equivalent $LCP(\bar{M}, \bar{q})$, where $\bar{q} = Eq$. Notice that given $\bar{x}$, a solution of $LCP(\bar{M}, \bar{q})$, the equivalent solution for $LCP(M, q)$ may be simply calculated as

$$x = D\bar{x}.$$

The classes of P_0 and P-matrices are invariant under scaling of the above type.

For our computational experiments we adopted a version of the EMD scaling known as balanced scaling [21]. Let

$$\begin{aligned} r(i) &= \max_j |M_{i,j}|, i = 1, \ldots, n, \\ M' &= R^{-1}M, \text{ where } R = \operatorname{diag}(r), \\ c(i) &= \max_i |M'_{i,j}|, j = 1, \ldots, n, \end{aligned}$$

and

$$\bar{M} = R^{-1}MC^{-1} \text{ where } C = \operatorname{diag}(c), \text{ and } \bar{q} = R^{-1}q.$$

If $\bar{x}^*, \bar{y}^*$ is an optimal solution of $LCP(\bar{M}, \bar{q})$, then $x^* = C^{-1}\bar{x}^*$ and $y^* = R\bar{y}^*$ is an optimal solution of $LCP(M, q)$. Once again consider the above simple example. After the balanced scaling, M becomes

$$\bar{M} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

We see that α disappears in $\bar{M}$. In fact, $\bar{M}$ is positive definite now. We will discuss some theoretical aspects of matrix scaling in §4.

For P-matrix LCPs, we generate matrix M and choose a complementary solution (x^*, y^*) such that $0 \leq x_i^*, y_i^* \leq 10$, $i = 1, \ldots, n$, $x^{*T}y^* = 0$, and $q = y^* - Mx^*$. The initial solutions are chosen, respectively, as $x^* + e$ and $y^* + Me$ (note that every entry of M is nonnegative). Table 3 shows the computational results with a P-matrix LCP for $n = 100$ and various values of α. It is clear that our scaling technique is effective for our potential reduction algorithm.

TABLE 3
P-matrix ($n = 100$).

	Without Scaling		With Scaling	
α	Avg. Itr.	Avg. CPU	Avg. Itr.	Avg. CPU
50	55.0	1.66	36.2	1.09
100	92.8	2.81	39.4	1.19
150	110.8	3.34	43.6	1.31
200	105.4	3.18	43.6	1.31
250	131.0	3.94	42.0	1.27
300	128.2	3.86	38.6	1.16

Table 4 summarizes the computational results with a P-matrix LCP using the scaling technique. The parameter α was chosen to be 50. Note that the average number of iterations for $\alpha = 50$ (with scaling, $n = 100$) in Table 3 is 36.2 and $\alpha = 50$ (with scaling, $n = 100$) in Table 4 is 34.6, respectively, since different random seeds were used in each experiment. It was observed that for an instance of a P-matrix LCP ($n = 100$ and $\alpha = 50$), 39 iterations are needed to solve the problem with $\beta = 0.99$ and the potential function values are strictly decreasing. On the other hand, it takes 65 iterations to solve it with $\beta = 0.5$.

TABLE 4
P-matrix ($\alpha = 50$).

n	Avg. Itr.	Avg. CPU
100	34.6	1.03
200	85.8	15.49
300	133.0	77.17
400	167.0	228.03
500	197.6	506.38

Table 5 illustrates the point that when the sizes of P_1 and P_2 are roughly equal, the problem becomes more difficult to solve. In this case the problem size n is 100 and the parameter α is chosen to be 50.

From these computational results, we observe the following:

1. As α is allowed to increase, the condition number of the matrix decreases. As the condition number worsens, the algorithm's convergence is slowed. This is exactly what we expected, based on our theoretical development.

TABLE 5
P-matrix with various n_1 and n_2 ($n = 100, \alpha = 50$).

n_1	n_2	Avg. Itr.	Avg. CPU
10	90	24.0	0.80
20	80	28.8	0.91
40	60	34.6	1.04
50	50	35.0	1.04
60	40	35.2	1.05
80	20	23.6	0.75
90	10	20.4	0.68

2. The row and column scaling technique, in general, helps to balance the matrix, and improve the convergence speed. This topic certainly has the potential for fruitful future research.

3. The convergence speed is sensitive to both the size and structure of the problem.

3.3. P_0-matrices. In many practical applications of LCPs, the M matrix is a P_0-matrix. As it was recently shown by Kojima et al. [10], a special P_0-LCP is NP-complete. Consider the matrix

$$M = \begin{pmatrix} P + D_1 & P + D_2 \\ -I_{\frac{n}{2}\times\frac{n}{2}} & 0 \end{pmatrix},$$

where $P \in R^{\frac{n}{2}\times\frac{n}{2}}$ is a positive definite matrix and D_1 and D_2 are diagonal matrices whose diagonal elements are positive. This problem has an engineering application considered by Kaneko [9].

We constructed the following test problems:

1. $P = A^T A$, where $A = \{a_{ij}\} \in R^{\frac{n}{2}\times\frac{n}{2}}$ and $0 < a_{ij} < 1$, $1 \leq (D_1)_{ii}, (D_2)_{ii} \leq 3$.
2. Let $x = \binom{x_1}{x_2}$ and $y = \binom{y_1}{y_2}$, where $x_1, y_1, x_2, y_2 \in R^{\frac{n}{2}}$. Generate x_1^*, y_2^* such that $0 \leq (x_1^*)_i, (y_2^*)_i < 10$ and $(x_1^*)_i + (y_2^*)_i > 0$, where $i = 1, \ldots, n/2$.
3. Generate y_1^* complement to x_1^* and x_2^* complement to y_2^*, such that $0 \leq (x_2^*)_i, (y_1^*)_i < 10$, where $i = 1, \ldots, n/2$.
4. Compute $q = y^* - Mx^* = \binom{q_1}{q_2}$, where $q_1, q_2 \in R^{\frac{n}{2}}$. Note that $q_2 > 0$.
5. $x_1^0 = q_2/2$ and $x_2^0 = G^{-1}c$, where $G = \operatorname{diag}((P + D_2)_{ii})$, and $c_i = |(q_1)_i|$, where $i = 1, \ldots, n/2$ and $y^0 = Mx^0 + q$. Then we can easily verify that $y^0 = Mx^0 + q > 0$.

From the problem's construction, we clearly see that the feasible region Ω has nonempty interior. We further show that the conditions of Proposition 4 are met in this problem.

PROPOSITION 5. *Let M and q be generated as above. Then, M is a P_0-matrix and the set $\{(x, y) \in \Omega^+ : \phi(x, y) \leq O(\rho L)\}$ is bounded.*

Proof. M being a P_0-matrix is due to Kaneko [9]. From (2.7), it is sufficient to prove that $\{(x, y) \in \Omega^+ : x^T y \leq \exp(L)\}$ is bounded. First, we see that x_1 (y_2) is bounded in Ω: $0 \leq x_1 \leq q_2$. Now it is sufficient to prove that x_2 is bounded. Note that

$$(x_1)^T y_1 + (x_2)^T y_2 = (x_1)^T (P + D_1) x_1 + (x_1)^T P x_2 + (x_1)^T D_2 x_2 + (x_1)^T q_1 + (x_2)^T (q_2 - x_1).$$

Since x_1 is bounded, $x_1 \geq 0$, $x_2 \geq 0$, and every component of P is nonnegative, $(x_1)^T (P + D_1) x_1 + (x_1)^T P x_2 + (x_1)^T q_1$ is bounded from below. Thus, $(x_1)^T D_2 x_2 + (x_2)^T (q_2 - x_1)$ is bounded from above. Therefore, x_2 must be bounded in Ω. □

Table 6 shows the computational results with P_0-matrix LCP test problems described above using the balanced scaling technique. For an instance of these test problems with $n = 100$, the algorithm needs 277 iterations to solve the problem with $\beta = 0.99$ and 449 iterations with $\beta = 0.5$. But, for instances of NP-complete problems introduced in [10] (P_0-matrix LCP), it was observed that β must be very small (e.g., 0.01 for $n = 31$) in order to guarantee the monotonic decrease of the potential function value. From the above observation, we conclude that a line search may be needed to minimize the potential function along the direction p^k.

TABLE 6
P_0-matrix.

n	Avg. Itr.	Avg. CPU
100	248.6	6.3
200	569.2	83.7
300	704.4	308.0

From these computational results, we see that:

1. The algorithm does converge, which is encouraging.
2. However, the potential reduction algorithm is currently far from being an adequate technique for solving the P_0-matrix LCP.

4. Matrix scaling. We have previously used matrix scaling to improve practical efficiency for solving P-matrix LCPs. For any n by n P-matrix M, a theoretical question is whether there exist scaling matrices E and D such that EMD is positive definite (PD). It can be verified that if $n = 2$, the answer is "yes." However, for $n \geq 3$ the answer is "no." For example,

$$M = \begin{pmatrix} 2 & 4 & 0 \\ 0 & 1 & 2 \\ 4 & 0 & 1 \end{pmatrix}$$

cannot be scaled into a PD matrix.

Since

$$EMD + DM^TE = E(MDE^{-1} + E^{-1}DM^T)E,$$

we see that the definiteness of $EMD + DM^TE$ depends only on the definiteness of $MDE^{-1} + E^{-1}DM^T$. Without loss of generality, we only use column scaling, i.e., MD, in our subsequent discussion. We now prove the following theorem.

PROPOSITION 6. *Let P_1 and P_2 be positive definite matrices, and let*

$$M = \begin{pmatrix} P_1 & Q \\ 0 & P_2 \end{pmatrix}.$$

Then, M is a P-matrix for any matrix Q. Furthermore, there exists a scalar $\beta > 0$ such that

$$\bar{M} = \begin{pmatrix} P_1 & \beta Q \\ 0 & \beta P_2 \end{pmatrix}$$

is PD.

Proof. Note

$$\bar{M} + \bar{M}^T = \begin{pmatrix} P_1 + P_1^T & \beta Q \\ \beta Q^T & \beta(P_2 + P_2^T) \end{pmatrix}.$$

Let $P = P_1 + P_1^T$ and $S = P_2 + P_2^T$. Then, $\bar{M} + \bar{M}^T$ can be decomposed as

$$\begin{pmatrix} I & 0 \\ \beta Q^T P^{-1} & I \end{pmatrix} \begin{pmatrix} P & 0 \\ 0 & \beta S - \beta^2 Q^T P^{-1} Q \end{pmatrix} \begin{pmatrix} I & \beta P^{-1} Q \\ 0 & I \end{pmatrix}.$$

Now the question is if there exists a $\beta > 0$ such that $\beta S - \beta^2 Q^T P^{-1} Q$ is PD. The answer is "yes" since S is PD and $Q^T P^{-1} Q$ is fixed, so that $\beta(S - \beta Q^T P^{-1} Q)$ is PD for a small enough positive β by continuity. □

Proposition 6 says that some P-matrices can be scaled into PD matrices. These include the triangular P-matrices, block triangular matrices where every diagonal block is PD, or block triangular P-matrices where every diagonal block is at most 2 by 2. This property may not have algorithmic value since triangular P-matrix LCPs can be solved by back substitution. However, we hope it partially explains why some matrix scaling can improve the practical efficiency for solving P-matrix LCPs as indicated in our computational results.

5. Concluding remarks. We have presented our computational experience with classes of PSD-matrix, P-matrix, and P_0-matrix LCPs. These test problems include the worst case examples of Murty and Fathi, randomly generated positive semi-definite LCPs, and various LCPs arising from engineering problems.

Based on these preliminary experiments, we make the following remarks:

1. The interior point algorithm has the potential for solving large-scale LCPs. As we expected, our algorithm did converge for all test problems we tried.

2. The convergence speed worsens from PSD, through P- to P_0-matrix LCPs. This reinforces our theoretical development. Significant difficulties exist in solving P_0-matrix LCPs. Pivoting algorithms are possibly more efficient for some P-matrix problems, and for some P_0-matrix problems when they work.

3. Although the matrix scaling technique may improve the theoretical or practical efficiency, there exists an obvious dependency between the dimension of the problem and the number of iterations required. Other row and column scaling techniques warrant further study and experimentation.

4. In our current experiment, the step size was heuristically taken as 99% of the way to the boundary. This strategy works for the PSD-matrix LCPs, but may not work for all P_0-matrix LCPs. A line search to minimize the potential function along the direction p^k may be needed. We plan to experiment with that approach.

5. To be really useful in practice, parallel implementation is absolutely necessary to efficiently solve each iteration of the interior point algorithm. If the real problems are sparse, the computational time is significantly improved according to our experience with convex quadratic programming problems [8].

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative matrices in the mathematical sciences*, Academic Press, New York, 1979.

[2] R. W. Cottle and G. B. Dantzig, *Complementary pivot theory in mathematical programming*, Linear Algebra Appl., 1 (1968), pp. 103–125.

[3] B. C. Eaves, *On the basic theorem of complementarity*, Math. Programming, 1 (1971), pp. 68–75.

[4] C. M. Elliot and J. R. Ockendon, *Weak variational methods for moving boundary problems*, in Research notes in Mathematics, Pitman, Boston, 1982.

[5] Y. Fathi, *Computational complexity of* LCPs *associated with positive definite matrices*, Math. Programming, 17 (1979), pp. 335–344.

[6] M. FIEDLER, *Special matrices and their applications in numerical mathematics*, Martinus Nijhoff Publishers, Dordrech, the Netherlands, 1986.

[7] M. FIEDLER AND V. PTÁK, *On matrices with nonpositive off-diagonal elements and positive principal minors*, Czechoslovak Math. J., 12 (1962), pp. 382–400.

[8] C. G. HAN, P. M. PARDALOS, AND Y. YE, *Computational aspects of an interior point algorithm for quadratic programming problems with box constraints*, in Large-Scale Numerical Optimization, T. Coleman and Y. Li, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990, pp. 92–112.

[9] I. KANEKO, *A linear complementarity problem with an n by 2n "P"-matrix*, Math. Programming Study, 7 (1978), pp. 120–141.

[10] M. KOJIMA, N. MEGIDDO, T. NOMA, AND A. YOSHISE, *A unified approach to interior point algorithms for linear complementarity problems*, Research Report RJ 7493, IBM Almaden Research Center, San Jose, CA, 1990.

[11] M. KOJIMA, N. MEGIDDO, AND Y. YE, *An interior point potential reduction algorithm for the linear complementarity problem*, Research Report RJ 6486, IBM Almaden Research Center, San Jose, CA, 1988.

[12] O. L. MANGASARIAN, *Characterization of linear complementarity problems as linear programs*, Math. Programming Study, 7 (1978), pp. 74–87.

[13] R. MATHIAS, *An improved bound for a fundamental constant associated with a P-matrix*, Appl. Math. Lett., 2 (1989), pp. 297–300.

[14] N. MEGIDDO, *A note on the complexity of P-matrix LCP and computing an equilibrium*, Report RJ6439, IBM Almaden Research Center, San Jose, CA, 1988.

[15] K. G. MURTY, *Computational complexity of complementary pivot methods*, Math. Programming Study, 7 (1978), pp. 61–73.

[16] ———, *Linear Complementarity, Linear and Nonlinear Programming*, Heldermann, Verlag, Berlin, 1988.

[17] P. M. PARDALOS AND J. B. ROSEN, *Global optimization approach to the linear complementarity problem*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 341–353.

[18] P. M. PARDALOS, *Linear complementarity problems solvable by integer programming*, Optimization, 19 (1988), pp. 467–474.

[19] D. SOLOW, R. STONE, AND C. A. TOVEY, *Solving* LCP *on P-matrix is probably not* NP*-hard*, Georgia Institute of Technology, Atlanta, manuscript, 1987.

[20] M. J. TODD AND Y. YE, *A centered projective algorithm for linear programming*, Math. Oper. Res., 15 (1990), pp. 508–529.

[21] J. A. TOMLIN, *On scaling linear programming problems*, Math. Programming Study, 4 (1975), pp. 146–166.

[22] Y. YE, *A further result on the potential reduction algorithm of the P-matrix linear complementarity problem*, Advances in Optimization and Parallel Computing, P. M. Pardalos, ed., North-Holland, Amsterdam, New York, 1992, pp. 310–316.

[23] ———, *An $O(n^3L)$ potential reduction algorithm for linear programming*, Math. Programming, 50 (1991), pp. 239–258.

[24] Y. YE AND P. M. PARDALOS, *A class of linear complementarity problems solvable in polynomial time*, Linear Algebra Appl., 152 (1991), pp. 3–17.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1061–1063, October 1993

010

APPROXIMATION OF MATRIX-VALUED FUNCTIONS*

ROY MATHIAS[†]

Abstract. It is shown that if $f(z) = \sum_{i=0}^{\infty} a_i z^i$ (the series having radius of convergence R) then for any $n \times n$ matrix A with spectral radius less than R and any norm $\|\cdot\|$ on the space of $n \times n$ matrices

$$\left\| f(A) - \sum_{i=0}^{k} a_i A^i \right\| \le \frac{1}{(k+1)!} \max_{s\in[0,1]} \|A^{k+1} f^{(k+1)}(sA)\|.$$

It is also shown that expressions for the error in numerical integration rules can be generalized to matrix-valued functions. The main point of the paper is that in these two cases it is not necessary to increase the bound by a factor depending on n when generalizing an inequality for scalar-valued functions to $n \times n$ matrix-valued functions.

Key words. matrix-valued function, truncated Taylor series, numerical integration

AMS subject classifications. 65D30, 15A45

Let M_n denote the space of $n \times n$ complex matrices. Let $\rho(\cdot)$ denote the spectral radius and let $\mathrm{tr}(\cdot)$ denote the trace on M_n. In this paper we show how some bounds for scalar functions can be generalized to norm bounds for matrix-valued functions without introducing a factor depending on n (the size of the matrices). The technique depends on using the dual norm to reduce the problem to the scalar case.

Given a norm $\|\cdot\|$ on M_n, we define its dual (with respect to the inner product $\mathrm{tr}AB^*$) by

$$\|A\|^D \equiv \max\{\mathrm{Re}[\mathrm{tr}(AB^*)] : \|B\| \le 1\}.$$

It follows from this definition that for any $A, B \in M_n$

$$|\mathrm{Re}[\mathrm{tr}(AB^*)]| \le \|A\| \, \|B\|^D. \tag{1}$$

The duality theorem for norms (see e.g., [2, Thm. 5.5.14]) states that $(\|\cdot\|^D)^D = \|\cdot\|$ and, consequently,

$$\|A\| = \max\{\mathrm{Re}[\mathrm{tr}(AB^*)] : \|B\|^D \le 1\}. \tag{2}$$

Our first result deals with the approximation of $f(A)$ by its truncated Taylor series. Corollary 2 improves the bound in [1, Thm. 11.2.4] by a factor of n in the case of the spectral norm.

THEOREM 1. *Let $f : [0,1] \to M_n$ be $k+1$ times continuously differentiable. Then for any norm $\|\cdot\|$ on M_n,*

$$\left\| f(1) - \sum_{i=0}^{k} \frac{f^{(i)}(0)}{i!} \right\| \le \frac{1}{(k+1)!} \max_{s\in[0,1]} \|f^{(k+1)}(s)\|. \tag{3}$$

[*] Received by the editors October 15, 1991; accepted for publication (in revised form) January 31, 1992.

[†] Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23187 (na.mathias@na-net.ornl.gov). This research was supported by a Summer Research grant from the College of William and Mary.

When we say that f is $k+1$ times continuously differentiable we mean that each of the functions $g_{ij}(t) = (f(t))_{ij}, i, j = 1, \ldots, n$, is a $k+1$ times continuously differentiable function of the real variable t.

Proof. Let f and k satisfy the conditions of the theorem. Let $B \in M_n$ be such that $\|B\|^D \leq 1$ and

$$\left|\operatorname{tr}\left[f(1) - \sum_{i=0}^{k} \frac{f^{(i)}(0)}{i!}\right] B^*\right| = \left\|f(1) - \sum_{i=0}^{k} \frac{f^{(i)}(0)}{i!}\right\|.$$

The existence of such a B is guaranteed by (2). Define $g(t) = \operatorname{Re}[\operatorname{tr}(f(tA)B^*)]$. Then

$$g^{(i)}(t) = \operatorname{Re}[\operatorname{tr}(f^{(i)}(t)B^*)]. \tag{4}$$

So now, by the formula for the error in the truncated Taylor series for a real-valued function we have

$$\begin{aligned}\left\|f(1) - \sum_{i=0}^{k} \frac{f^{(i)}(0)}{i!}\right\| &= \left|\operatorname{tr}\left[f(A) - \sum_{i=0}^{k} \frac{f^{(i)}(0)}{i!}\right] B^*\right| \\ &= \left|g(1) - \sum_{i=0}^{k} \frac{g^{(i)}(0)}{i!}\right| \\ &\leq \frac{1}{(k+1)!} \max_{s\in[0,1]} |g^{(k+1)}(s)| \\ &= \frac{1}{(k+1)!} \max_{s\in[0,1]} |\operatorname{Re}[\operatorname{tr}(f^{(k+1)}(s)B^*)]| \\ &\leq \frac{1}{(k+1)!} \max_{s\in[0,1]} \|f^{(k+1)}(s)\|.\end{aligned}$$

We have used (4) for the third equality and (1) for the final inequality. □

COROLLARY 2. *Let g be given by the series*

$$g(z) = \sum_{i=0}^{\infty} a_i z^i.$$

Let R be the radius of convergence of this series. Let $A \in M_n$ be given with $\rho(A) < R$. Then for any norm $\|\cdot\|$ on M_n and $k = 0, 1, \ldots$

$$\left\|g(A) - \sum_{i=0}^{k} a_i A^i\right\| \leq \frac{1}{(k+1)!} \max_{s\in[0,1]} \|A^{k+1} g^{(k+1)}(sA)\|. \tag{5}$$

Proof. Take $f(t) = g(tA)$, check that

$$f^{(i)}(t) = A^i g^{(i)}(tA),$$

and apply Theorem 1. □

Corollary 2 can also be proven by showing

$$g(A) = \sum_{i=0}^{k} a_i A^i + \frac{1}{k!} \int_0^1 t^k A^{k+1} f^{(k+1)}(tA) dt \tag{6}$$

(using, for example, the method outlined in [3, pp. 444–445]) and then taking norms.

Our next result deals with the error in approximating the integral of a matrix-valued function by a sum. In the case of the spectral norm it improves on the bound in [1, Eq. 11.2.7] by a factor of n. See [4] for an application of this bound. We omit the proof of this result as it can be reduced to the scalar case by the same technique as was used in the proof of Theorem 1.

THEOREM 3. *Let $c > 0$, $x_i \in [0,1]$ and $a_i \in R$, $i = 1, 2, \ldots, m$ be such that for any function g that is $k+1$ times differentiable on* $[0,1]$

$$\left| \int_0^1 g(x)dx - \sum_{i=1}^{m} a_i g(x_i) \right| \leq c \max_{s \in [0,1]} |g^{(k+1)}(s)|. \tag{7}$$

Let $f : [0,1] \rightarrow M_n$ be such that $f^{(k+1)}(t)$ exists for $t \in (0,1)$. Then for any norm $\|\cdot\|$ on M_n

$$\left\| \int_0^1 f(x)dx - \sum_{i=1}^{m} a_i f(x_i) \right\| \leq c \max_{s \in [0,1]} \|f^{(k+1)}(s)\|. \tag{8}$$

Note that in Theorem 3 we have not assumed that f is of any special form, but merely that it is sufficiently differentiable. For example, one can fix $X, Z \in M_n$ and take $f(t) = e^{-tX} Z e^{tX}$ as in [4]. Alternatively, one could fix $X \in M_n$ and define the i, j element of $f(t)$ by $f(t)_{ij} = e^{tx_{ij}}$.

Acknowledgment. I am grateful to Professor R. Horn for pointing out (6).

REFERENCES

[1] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, second edition, 1989.

[2] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[3] ———, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[4] R. Mathias, *Evaluating the Frechet derivative of the matrix exponential*, Numer. Math., 63 (1992), pp. 213–226.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1064–1083, October 1993

011

A DECOMPOSITION FOR THREE-WAY ARRAYS*

S. E. LEURGANS[†], R. T. ROSS[‡], AND R. B. ABEL[§]

Abstract. An I-by-J-by-K array has rank 1 if the array is the outer product of an I-, a J-, and a K-vector. The authors prove that a three-way array can be uniquely decomposed as the sum of F rank-1 arrays if the F vectors corresponding to two of the ways are linearly independent and the F vectors corresponding to the third way have the property that no two are collinear. Several algorithms that implement the decomposition are described. The algorithms are applied to obtain initial values for nonlinear least-squares calculations. The performances of the decompositions and of the nonlinear least-squares solutions on real and on simulated data are compared. An extension to higher-way arrays is introduced, and the method is compared with those of other authors.

Key words. alternating least-squares algorithm, array rank, multiway arrays

AMS subject classifications. 15A23, 15A69, 62H25, 62J99

1. Introduction. This paper gives a method for identifying the parameters of a multilinear model proposed by Harshman [11] and by Carroll and Chang [4]. The models, described in §3 below, represent three-way arrays in terms of three sets of parameters, one set associated with each way. To work with these models, it is necessary to be able to identify the parameters for each way from the three-way array. Kruskal [14] gives conditions on the sets of parameters such that the parameters are identifiable, but his paper does not include a construction. Assuming conditions stronger than his weakest conditions, we show how to recover the parameters from the array. This method is related to the algebraic initialization method of Sands and Young [25]. Our method requires one eigenvector decomposition, one Moore–Penrose generalized inverse, and a finite number of arithmetic steps. These models have been applied recently in the chemical literature, as described in §7 below. In this paper we concentrate on a mathematical presentation to make these ideas more widely applicable. Using Moore–Penrose inverses gives a natural derivation, and we provide some evidence on the effectiveness of an appropriate decomposition for initializing alternating least-squares iterations.

Section 2 reviews some facts about matrices, or two-way arrays. The models for three-way arrays of interest to us are outlined in §3. We require a decomposition that returns the parameters ($\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{\Gamma}$ in the notation of §3) from a three-way array $\boldsymbol{\mu}$. Section 3 gives identifiability conditions under which a theorem is proven in §4. The theorem characterizes the parameters of the model. Section 5 shows how the theorem can be applied to give an algorithm. The section also reports some of our experience in applying this decomposition to initialize alternating least-squares calculations. We summarize the behavior of the decompositions and of the least-squares calculations for

* Received by the editors September 10, 1990; accepted for publication (in revised form) January 1, 1992. This research was supported by National Science Foundation grants DMS-88-05402 and DMS-89-02265 at Ohio State University and by the Ohio Supercomputer Center. Revision of the manuscript was supported by National Science Foundation grant HRD-9103314 while the first author was visiting the University of Chicago.

[†] Department of Statistics, Ohio State University, Columbus, Ohio 43210-1247 (leurgans@is.rpslmc.edu).

[‡] Department of Biochemistry, Ohio State University, Columbus, Ohio 43210-1229 (rtr+@ohio-state.edu).

[§] National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892-9905.

five sets of real data and a few sets of artificial data. In §6 we discuss decompositions for higher-way arrays. Section 7 compares our work with other papers.

The models of §3 can be interpreted as a representation of an array as the sum of rank-1 arrays. Since rank-1 arrays are often easy to interpret, these models can lead to simple interpretations of the array. For two-way arrays, or matrices, such representations are never unique without side conditions that often interfere with the simplicity of the interpretation. Frequently the condition that the vectors for each way be an orthogonal set is imposed. Since negative parameters have no physical meaning in many applications of these models, orthogonality cannot always provide representations that are interpretable unless the array is a rank-1 array. However, if the original array has more than two ways, the array often has a unique representation as the sum of rank-1 arrays. As has been emphasized in the literature on these models, side conditions, such as orthogonality, are not required. The sufficient conditions of §3 permit representations that do not sacrifice interpretability. Since representations for three-way arrays can lead to interpretable parameters, experiments that use three independent variables to generate three ways of an array can remove ambiguities unavoidable in experiments that use only two of the independent variables.

2. Decompositions of matrices. The decompositions for three-way arrays rely on results for two-way arrays, or matrices. The results we will use have been collected in this section. We state one proposition concerning the Moore–Penrose generalized inverse of products of matrices. This proposition implies that the Moore–Penrose generalized inverse can be obtained from a singular value decomposition.

A matrix has rank 1 if it is the outer product of one pair of vectors. An I-by-J matrix $\boldsymbol{\mu}$ has rank F if F is the smallest integer such that $\boldsymbol{\mu}$ is a sum of rank-1 matrices, or

$$\boldsymbol{\mu} = \sum_{f=1}^{F} \boldsymbol{\alpha}_f \times \boldsymbol{\beta}_f = \boldsymbol{A}\boldsymbol{B}^T, \tag{1}$$

where the columns of $\boldsymbol{A}$ are $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_F$ and those of $\boldsymbol{B}$ are $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_F$ and $\times$ denotes the outer product defined by $\boldsymbol{\alpha}_f \times \boldsymbol{\beta}_f$, the I-by-J matrix whose i, j entry is $\alpha_f[i]\beta_f[j]$. The model (1) is a bilinear model for $\boldsymbol{\mu}$. To avoid trivial reparameterizations we require the columns of $\boldsymbol{A}$ to be unit vectors whose first nonzero element is positive. When $F = 1$, the model is identifiable up to the sign convention: $\boldsymbol{\alpha}$ is any column of $\boldsymbol{\mu}$ divided by its length and $\boldsymbol{\beta}$ is the vector of the lengths of the columns, where the jth component is multiplied by the sign of the first nonzero element of the column. When $F > 1$, the parameters are not identifiable because $\boldsymbol{AQ}(\boldsymbol{BQ})^T = \boldsymbol{AB}^T$ when $\boldsymbol{Q}$ is any orthogonal F-by-F matrix preserving the sign convention for $\boldsymbol{A}$. This nonidentifiability leads to the well-known rotation problem in factor analysis. If the columns of $\boldsymbol{A}$ and $\boldsymbol{B}$ are forced to be orthonormal, then the singular-value decomposition is the unique (up to permutation) decomposition of $\boldsymbol{\mu}$, provided that the singular values are unique [12, p. 414]. The singular-value decomposition of an I-by-J matrix $\boldsymbol{M}$ with rank F is an I-by-F matrix $\boldsymbol{U}$ with orthonormal columns, a J-by-F matrix $\boldsymbol{V}$ with orthonormal columns, and an F-by-F diagonal matrix $\boldsymbol{S}$ with nonincreasing positive diagonal elements satisfying $\boldsymbol{M} = \boldsymbol{USV}^T$. The diagonal elements of $\boldsymbol{S}$ are the singular values of $\boldsymbol{M}$. We follow Greenacre [8] by taking singular values to be nonzero. The columns of $\boldsymbol{U}$ are the left singular vectors of $\boldsymbol{M}$, and the columns of $\boldsymbol{V}$ are the right singular vectors of $\boldsymbol{M}$. Mathematically, the requirement that the columns of $\boldsymbol{A}$ be mutually orthogonal is very natural, but practical interpretation is frequently compromised.

The Moore–Penrose generalized inverse is a natural extension of matrix inversion to matrices that may be singular or even rectangular. (For defining properties see, for example, [26], [8], [20].) We use $\boldsymbol{M}^+$ to denote the Moore–Penrose generalized inverse of the matrix $\boldsymbol{M}$. We state a property of Moore–Penrose generalized inverses as a proposition.

PROPOSITION 2.1. *If* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *have equal full column rank and if the square matrix* $\boldsymbol{D}$ *is nonsingular, then*

$$(\boldsymbol{A}\boldsymbol{D}\boldsymbol{B}^T)^+ = (\boldsymbol{B}^+)^T\boldsymbol{D}^{-1}\boldsymbol{A}^+. \tag{2}$$

This proposition follows from Greville [9] and is given as Corollary 5 of Theorem 2.16 of Pringle and Rayner [20, p. 31]. Direct verification follows from routine applications of the defining properties of Moore–Penrose generalized inverses. If the hypotheses fail, Greville's results show that (2) can be false.

3. PARAFAC models for three-way arrays. In this section we describe the models of interest. Interpretation of these models for three-way arrays is much easier when the parameters can be identified from the array. We give three identifiability conditions that are sufficient for identifiability if three parameterization conventions are also imposed. The parameterization conventions can (essentially) always be imposed whenever the model holds. The conventions remove some trivial nonidentifiabilities. The results of Kruskal [14], discussed after the identifiability conditions are introduced, guarantee that IC1, IC2, and IC3 are sufficient for identifiability of all the parameters. Kruskal also provides weaker conditions, although his paper does not construct the parameters from the array. Since §5 provides such a construction, the identifiability of the parameters is verified. This section concludes with a presentation of some properties of the models.

The same models for three-way arrays were used in methods proposed independently by Harshman [10] and by Carroll and Chang [4]. The former paper calls the models *parallel factor models*, or PARAFAC, and the latter paper refers to *canonical decompositions*, or CANDECOMP. Both acronyms correspond to computer programs. For recent discussions of the methods in their original contexts, see [15], [2], [7]. Appellof and Davidson [1] first applied these models to spectroscopy.

The F-factor PARAFAC model for an I-by-J-by-K three-way array $\boldsymbol{\mu}$ is that $\boldsymbol{\mu}$ satisfies the *trilinear model*:

$$\boldsymbol{\mu} = \sum_{f=1}^{F} \boldsymbol{\alpha}_f \times \boldsymbol{\beta}_f \times \boldsymbol{\gamma}_f. \tag{3}$$

Since each array $\boldsymbol{\alpha}_f \times \boldsymbol{\beta}_f \times \boldsymbol{\gamma}_f$ is a rank-1 array, an F-term trilinear model (3) holds if $\boldsymbol{\mu}$ is the sum of F rank-1 arrays [14]. Since the trilinear model is preserved if $\boldsymbol{\alpha}_f$, $\boldsymbol{\beta}_f$, and $\boldsymbol{\gamma}_f$ are multiplied by constants whose product is 1 and since the values of f can be permuted without changing the sum, some parameterization conventions are necessary to provide identifiability. Before introducing either the identifiability conditions or a set of parameterization conventions, we rewrite the trilinear model in terms of matrices.

The trilinear model (3) can be written in terms of the K I-by-J matrices $\boldsymbol{\mu}_k$ as follows:

$$\boldsymbol{\mu}_k = \boldsymbol{\mu}[,,k] = \sum_{f=1}^{F} \boldsymbol{\alpha}_f \times \boldsymbol{\beta}_f \gamma_f[k] = \sum_{f=1}^{F} \boldsymbol{\alpha}_f \gamma_f[k] \boldsymbol{\beta}_f^T = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{B}^T,$$

where $\boldsymbol{A}$ is a real matrix whose columns are the $\boldsymbol{\alpha}_f$'s, $\boldsymbol{B}$ is a real matrix whose columns are the $\boldsymbol{\beta}_f$'s, and $\boldsymbol{D}_k$ is the F-by-F diagonal matrix whose diagonal elements are $\gamma_f[k], f = 1, \ldots, F$. Kruskal [14] refers to the $\boldsymbol{\mu}_k$'s as 3-slabs of the array $\boldsymbol{\mu}$, and we refer to them as *slabs* below. Note that the diagonal of $\boldsymbol{D}_k$ is the kth row of $\boldsymbol{\Gamma}$. Therefore, the PARAFAC model for the array $\boldsymbol{\mu}$ is equivalent to requiring that the K matrices $\boldsymbol{\mu}_k$ can be factored into the product of an I-by-F, an F-by-F, and an F-by-J matrix in such a way that the K I-by-F matrices are identical, the K F-by-J matrices are identical, and the K F-by-F matrices are all diagonal. The diagonal matrices $\boldsymbol{D}_k$ vary with k and can be thought of as reflecting the relative importance of the $\boldsymbol{\alpha}_f$'s and the $\boldsymbol{\beta}_f$'s as k varies. The equation displayed above can be summarized as *trilinear matrix equations* (4):

$$(4) \qquad \boldsymbol{\mu}_k = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{B}^T, \qquad k = 1, \ldots, K.$$

The trivial nonidentifiabilities alluded to in the preceding paragraph correspond to modification of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{\Gamma}$ by multiplication on the right by diagonal matrices whose product is the identity matrix or by applying the same permutation to their columns.

The trivial sources of nonidentifiability are not the only ones: Conditions on $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{\Gamma}$ are required. One set of sufficient conditions is that all three matrices have the property that every pair of columns is linearly independent and that two of the matrices have all columns linearly independent. For specificity, we adopt the following *identifiability conditions*:

(IC1) The columns of $\boldsymbol{A}$ are linearly independent.
(IC2) The columns of $\boldsymbol{B}$ are linearly independent.
(IC3) Every pair of columns of $\boldsymbol{\Gamma}$ is linearly independent.

In §4, we show how to recover the parameters under these conditions.

Of course, when these three conditions hold, all F columns of $\boldsymbol{A}$ are linearly independent, all F columns of $\boldsymbol{B}$ are linearly independent, and every pair of columns of $\boldsymbol{\Gamma}$ are linearly independent. In Kruskal's notation, $I_o \geq F$, $J_o \geq F$, and $K_o \geq 2$, so that $I_o + J_o + K_o \geq 2F + 2$. This last inequality is sufficient for the parameters to be uniquely defined, by [14, Thm. 4a]. The same theorem implies that uniqueness can still hold when IC1 and IC2 fail if IC3 is strengthened. For example, if the last column of $\boldsymbol{A}$ is the sum of the preceding mutually linearly independent columns, if the last column of $\boldsymbol{B}$ is the sum of the preceding mutually linearly independent columns, and if every four columns of $\boldsymbol{\Gamma}$ are linearly independent, then Kruskal's results imply uniqueness, although our decomposition does not apply.

The identifiability conditions are also easy to interpret because each of the three conditions concerns the parameters associated with one way of the array. The first two identifiability conditions assert that there must truly be F factors present in two ways. The third condition requires that no two underlying factors respond to the other way in an exactly proportional manner. If $K > 1$, IC3 can hold when $F > K$. Observe that if $K = 1$, IC3 can hold only if $F = 1$. This observation corresponds to the nonidentifiability of bilinear models if $F > 1$. The conditions IC1 and IC2, however, can hold only if $F \leq I$ and $F \leq J$.

To see that condition IC3 cannot be omitted, suppose that $F = 2$ and that $\gamma_1 = c_1\gamma$ and $\gamma_2 = c_2\gamma$. Then the trilinear model implies that

$$\boldsymbol{\mu} = \boldsymbol{\alpha}_1 \times \boldsymbol{\beta}_1 \times c_1\gamma + \boldsymbol{\alpha}_2 \times \boldsymbol{\beta}_2 \times c_2\gamma = \{c_1\boldsymbol{\alpha}_1 \times \boldsymbol{\beta}_1 + c_2\boldsymbol{\alpha}_2 \times \boldsymbol{\beta}_2\} \times \gamma.$$

The I-by-J matrix in braces is

$$(c_1\boldsymbol{\alpha}_1 \vdots c_2\boldsymbol{\alpha}_2)(\boldsymbol{\beta}_1 \vdots \boldsymbol{\beta}_2)^T = (c_1\boldsymbol{\alpha}_1 \vdots c_2\boldsymbol{\alpha}_2)\boldsymbol{M}\boldsymbol{M}^{-1}(\boldsymbol{\beta}_1 \vdots \boldsymbol{\beta}_2)^T,$$

where $\boldsymbol{M}$ is any nonsingular 2-by-2 matrix. The arbitrariness of $\boldsymbol{M}$ implies that $\boldsymbol{A}$ and $\boldsymbol{B}$ cannot be identified. These calculations can be extended to show that failure of IC3 always prevents identifiability of the columns of $\boldsymbol{A}$ and of $\boldsymbol{B}$ corresponding to the collinear columns of $\boldsymbol{\Gamma}$.

The following *parameterization conventions* remove these trivial sources of nonidentifiability:

(PC1) The columns of $\boldsymbol{A}$ and of $\boldsymbol{B}$ are unit vectors.

(PC2) In each column of $\boldsymbol{A}$, the element with largest magnitude is positive. If several elements in a column have magnitude equal to the largest magnitude, then the first such element is positive.

(PC3) The diagonal elements of

$$\boldsymbol{D}_+ = \sum_{k=1}^{K} \boldsymbol{D}_k$$

are nonincreasing and positive, that is,

$$\mathbf{1}_K^T\gamma_1 = \boldsymbol{D}_+[1,1] \geq \cdots \geq \mathbf{1}_K^T\gamma_f = \boldsymbol{D}_+[f,f] \geq \cdots \mathbf{1}_K^T\gamma_F = \boldsymbol{D}_+[F,F] > 0.$$

The parameterization conventions are easy to interpret. Convention PC1 asserts that the constant is carried in the third way. Convention PC2 establishes a sign convention for $\boldsymbol{A}$ by assuming that the first element of the column achieving the maximum absolute value for the column is positive. The convention that the columns of $\boldsymbol{A}$ are all unit vectors implies that PC2 is not vacuous in that every column of $\boldsymbol{A}$ must have strictly positive maximum absolute value. Convention PC3 determines the signs of the γ_f's, leaving the sign of the $\boldsymbol{\beta}_f$'s determined, and it also fixes the ordering of the factors when the inequalities are strict. If the inequalities are not strict, additional conditions will be required for identifiability. The requirement that all of the elements of $\boldsymbol{D}_+$ be positive forces $\boldsymbol{\mu}$ to satisfy a genuinely F-factor model. If $\boldsymbol{\mu}$ satisfies a PARAFAC model with no more than $F-1$ factors, then $\boldsymbol{D}_+[F,F] = 0$, the Fth column of $\boldsymbol{\Gamma}$ is entirely 0, and the Fth columns of $\boldsymbol{A}$ and $\boldsymbol{B}$ cannot possibly be well defined. It is possible for PC3 to hold with $\boldsymbol{D}_+[F,F] = 0$ if two terms $\boldsymbol{\alpha}_f \times \boldsymbol{\beta}_f\gamma_f[k]$ and $\boldsymbol{\alpha}_{f'} \times \boldsymbol{\beta}_{f'}\gamma_{f'}[k]$ exactly cancel when summed over k. Since all elements of $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{\Gamma}$ are generally nonnegative in the models we study, we will not explicitly present a decomposition as valid when cancellations occur. However, it will become apparent in §4 that $\boldsymbol{D}_+$ can always be replaced in PC3 and in the decompositions by a matrix $\boldsymbol{D}(\boldsymbol{w})$ as defined there.

Once some of the parameters of the trilinear model are known, it is easy to determine the others. We first suppose that $\boldsymbol{A}$ is known, and we then show that IC1 implies that $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ can be determined from $\boldsymbol{\mu}$ and $\boldsymbol{A}$. Identifiability condition IC1 implies that $\boldsymbol{A}^T\boldsymbol{A}$ is nonsingular. It follows that $\boldsymbol{A}^+ = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ and that $\boldsymbol{A}^+\boldsymbol{A}$ is the F-dimensional identity matrix, or

$$\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{I}_F. \tag{5}$$

Equation (5) and the trilinear matrix equations (4) imply that

$$A^+\mu_k = A^+(AD_kB^T) = D_kB^T, k = 1, \ldots, K. \tag{6}$$

Summing (6) over k, we obtain

$$A^+\mu_+ = D_+B^T.$$

Since the rows of B^T have length 1 and D_+ is diagonal, the fth diagonal element of D_+ must be the length of the fth row of $A^+\mu_+$ and the fth column of B is just the unit vector formed by normalizing the fth row of $A^+\mu_+$. Algebraically, the diagonal of D_+ is the square root of the diagonal of $A^+\mu_+(A^+\mu_+)^T$ and

$$B = \mu_+^T(A^+)^T D_+^{-1}. \tag{7}$$

The diagonal matrices D_k, and hence Γ, can now be calculated from (6). We note that this argument does not require B to have full column rank.

If both A and B are known and satisfy IC1 and IC2, a convenient expression is available for D_k. Since $(B^T)^+ = (B^+)^T$, the derivation of (5) shows that IC2 implies that

$$B^T(B^T)^+ = I_F. \tag{8}$$

If both sides of (6) are multiplied on the right by $(B^T)^+$, it follows that

$$D_k = A^+\mu_k(B^T)^+, \qquad k = 1, \ldots, K. \tag{9}$$

4. Mathematical decompositions. This section contains decompositions that can be used to obtain A from μ when the identifiability conditions hold. Since (7) and (6) give B and D_k (and hence Γ) from μ and A, all of the parameters can be identified from μ if A can be determined from μ when neither B nor Γ is known. The first part of Theorem 4.1 states that identifiability conditions on the first two ways imply that the columns of A are eigenvectors of certain matrices computed from the array μ. The second part of Theorem 4.1 states that if a weaker identifiability condition for the third way also holds, the only vectors that are eigenvectors of all of the matrices are scalar multiples of the columns of A and vectors in the intersection of all of the kernels of the matrices. The theorem therefore determines A up to permutation and sign changes of the columns when all three identifiability conditions hold. The analogues of the theorem obtained by permuting ways of the array are left to the reader. The two corollaries reexpress the decompositions in forms that are useful in applications. Corollary 4.3 restates Theorem 4.1 after reduction to an F-by-F-by-K array η. Corollary 4.4 gives a variant appropriate when $I = J$ and all of the matrices $\mu[,,k]$ are symmetric.

The key to the decomposition is to compare two linear combinations of the matrices $\mu_k, k = 1, \ldots, K$. To that end, for every K-vector w, define

$$\mu(w) = \sum_{k=1}^{K} w[k]\mu[,,k]$$

and

$$D(w) = \sum_{k=1}^{K} w[k]D_k,$$

so that $\boldsymbol{\mu}_+ = \boldsymbol{\mu}(\mathbf{1}_K)$ and $\boldsymbol{D}_+ = \boldsymbol{D}(\mathbf{1}_K)$. Theorem 4.1 below presents an eigenvalue problem whose solution is the columns of $\boldsymbol{A}$. See [17] for an interpretation of the matrices $\boldsymbol{\theta}_k$ and a heuristic rationale for the theorem.

THEOREM 4.1. *Assume that the trilinear matrix equations* (4) *hold with the parameterization conventions* PC1–PC3. *Using the Moore–Penrose inverse of* $\boldsymbol{\mu}_+$, *we define* $\boldsymbol{\theta}_k = \boldsymbol{\mu}_k\boldsymbol{\mu}_+^+$ *and* $\boldsymbol{\theta}(\boldsymbol{w}) = \boldsymbol{\mu}(\boldsymbol{w})\boldsymbol{\mu}_+^+$.

1. *If* IC1 *and* IC2 *hold, then for every nonzero K-vector $\boldsymbol{w}$ the columns of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{\theta}(\boldsymbol{w})$ with eigenvalues equal to the diagonal elements of $\boldsymbol{D}(\boldsymbol{w})\boldsymbol{D}_+^{-1}$, $k = 1,\ldots,K$.*

2. *If identifiability conditions* IC1–IC3 *hold, then there exists a K-vector $\boldsymbol{w}$ such that the only nontrivial eigenvectors of $\boldsymbol{\theta}(\boldsymbol{w})$ are scalar multiples of the columns of $\boldsymbol{A}$.*

Proof. It suffices to establish the first claim for vectors $\boldsymbol{w}$ that are columns of the K-dimensional identity matrix or to show that every column of $\boldsymbol{A}$ is an eigenvector of each matrix $\boldsymbol{\theta}_k$, $k = 1,\ldots,K$. First we argue that the columns of $\boldsymbol{A}$ are indeed eigenvectors of the $\boldsymbol{\theta}_k$'s.

By IC1, IC2, and PC3, Proposition 2.1 can be applied to give the Moore–Penrose generalized inverse of $\boldsymbol{\mu}_+$ in terms of the generalized inverses of $\boldsymbol{A}$, $\boldsymbol{D}_+$, and $\boldsymbol{B}$:

$$\boldsymbol{\mu}_+^+ = (\boldsymbol{B}^T)^+\boldsymbol{D}_+^{-1}\boldsymbol{A}^+.$$

Substitution into the definition of $\boldsymbol{\theta}_k$ gives

$$\boldsymbol{\theta}_k = \boldsymbol{A}\boldsymbol{D}_k(\boldsymbol{B}^T)(\boldsymbol{B}^T)^+\boldsymbol{D}_+^{-1}\boldsymbol{A}^+.$$

By (8), $\boldsymbol{\theta}_k = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{D}_+^{-1}\boldsymbol{A}^+$, so that

$$\boldsymbol{\theta}_k\boldsymbol{A} = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{D}_+^{-1}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{D}_+^{-1}, \tag{10}$$

where the right-hand equality follows from (5). Since the matrix $\boldsymbol{D}_k{\boldsymbol{D}_+}^{-1}$ is diagonal, equation (10) demonstrates that the columns of $\boldsymbol{A}$ are right-eigenvectors of $\boldsymbol{\theta}_k$, with eigenvalues as claimed.

From part 1, each column of $\boldsymbol{A}$ is an eigenvector of $\boldsymbol{\theta}(\boldsymbol{w})$ for every $\boldsymbol{w}$. The second part of the theorem will follow if $\boldsymbol{\theta}(\boldsymbol{w})$ has F one-dimensional eigenspaces or if the matrix $\boldsymbol{\theta}(\boldsymbol{w})$ has F distinct real nonzero eigenvalues or if the diagonal elements of $\boldsymbol{D}(\boldsymbol{w})\boldsymbol{D}_+^{-1}$ are distinct and positive. Lemma 4.2 below shows that PC3 and IC3 imply that the diagonal elements of $\boldsymbol{D}(\boldsymbol{w})\boldsymbol{D}_+^{-1}$ are distinct and positive for almost all $\boldsymbol{w}$. The theorem follows. □

LEMMA 4.2. *If* PC3 *and* IC3 *hold, then the diagonal elements of $\boldsymbol{D}(\boldsymbol{w})\boldsymbol{D}_+^{-1}$ are distinct and positive unless $\boldsymbol{w}$ is in a set of Lebesgue measure* 0.

Proof. We first argue that the set of $\boldsymbol{w}$'s such that a diagonal element of $\boldsymbol{D}(\boldsymbol{w})\boldsymbol{D}_+^{-1}$ is zero has Lebesgue measure 0. By definition the fth diagonal element is 0 if and only if $\boldsymbol{w}^T\boldsymbol{\gamma}_f = 0$ or if $\boldsymbol{w}$ is in the orthogonal complement of $\boldsymbol{\gamma}_f$, a nonzero vector by PC3. The union of F orthogonal complements has Lebesgue measure 0, so that $\boldsymbol{D}(\boldsymbol{w})\boldsymbol{D}_+^{-1}$ has positive diagonal elements for almost all $\boldsymbol{w}$'s.

Next we show that if PC3 is assumed to hold, then if the fth and f'th diagonal elements of $\boldsymbol{D}(\boldsymbol{w})\boldsymbol{D}_+^{-1}$ are tied, then either IC3 fails or $\boldsymbol{w}$ is in a set of measure 0. The fth diagonal element of $\boldsymbol{D}(\boldsymbol{w})\boldsymbol{D}_+^{-1}$ is $\boldsymbol{w}^T\boldsymbol{\gamma}_f/\mathbf{1}_K^T\boldsymbol{\gamma}_f$. Therefore, the fth and f'th diagonal elements of $\boldsymbol{D}(\boldsymbol{w})\boldsymbol{D}_+^{-1}$ are equal if

$$\boldsymbol{w}^T\left(\frac{1}{\mathbf{1}_K^T\boldsymbol{\gamma}_f}\boldsymbol{\gamma}_f - \frac{1}{\mathbf{1}_K^T\boldsymbol{\gamma}_{f'}}\boldsymbol{\gamma}_{f'}\right) = 0, f \neq f'. \tag{11}$$

If the vector expression in parentheses above is zero,

$$\gamma_{f'} = \frac{\mathbf{1}_K^T \gamma_{f'}}{\mathbf{1}_K^T \gamma_f} \gamma_f.$$

Since PC3 implies that the coefficient of γ_f above is positive and finite, γ'_f is a scalar multiple of γ_f, or the columns f and f' of $\boldsymbol{\Gamma}$ are collinear, so that IC3 fails.

Thus if IC3 holds, none of the vectors in parentheses in (11) can be 0, or

$$\frac{1}{\mathbf{1}_K^T \gamma_f} \gamma_f - \frac{1}{\mathbf{1}_K^T \gamma_{f'}} \boldsymbol{\Gamma}_{f'} \neq 0, f' \neq f, f = 1, \ldots, F; f' = 1, \ldots, F.$$

That is, for a pair of diagonal elements of $\boldsymbol{D}(\boldsymbol{w})\boldsymbol{D}_+^{-1}$ to be tied, equation (11) implies that $\boldsymbol{w}$ must be in the orthogonal complement of one of the nonzero K-vectors above. Since each orthogonal complement is a hyperplane of dimension $K-1$, the diagonal elements will be distinct unless $\boldsymbol{w}$ lies in the union of $F(F-1)/2$ hyperplanes of dimension $K-1$. The lemma follows. □

When I (or J) is much larger than F, reparameterization can remove redundant parameters and reduce the order of the matrices whose eigenvalues are sought. Using PC3, IC1, and IC2, we see that the column space of $\boldsymbol{\mu}_+$ is the same as the column space of $\boldsymbol{A}$ and that the column space of each $\boldsymbol{\mu}_k$ is a subspace of the column space of $\boldsymbol{A}$. Therefore, if $\boldsymbol{U}$ is an I-by-F matrix whose orthonormal columns span the column space of $\boldsymbol{\mu}_+$ and if $\boldsymbol{A}' = \boldsymbol{U}^T\boldsymbol{A}$, then $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{A}'$. Therefore, $\boldsymbol{A}$ can be determined by determining the F-by-F matrix $\boldsymbol{A}'$. Similarly, if the orthonormal columns of $\boldsymbol{V}$ span the column space of $\boldsymbol{\mu}_+^T$, setting $\boldsymbol{B}' = \boldsymbol{V}^T\boldsymbol{B}$ implies that $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{B}'$. Substituting for $\boldsymbol{A}$ and $\boldsymbol{B}$ in the trilinear matrix equations (4) gives $\boldsymbol{\mu}_k = \boldsymbol{U}\boldsymbol{A}'\boldsymbol{D}_k\boldsymbol{B}'^T\boldsymbol{V}^T$. Since both $\boldsymbol{U}$ and $\boldsymbol{V}$ have orthonormal columns,

$$\boldsymbol{\mu}'_k = \boldsymbol{U}^T\boldsymbol{\mu}_k\boldsymbol{V} = \boldsymbol{A}'\boldsymbol{D}_k\boldsymbol{B}'^T, \qquad k = 1, \ldots, K,$$

are K F-by-F matrices satisfying an F-factor PARAFAC model. Convenient matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are those obtained from the singular-value decomposition of $\boldsymbol{\mu}_+ = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$, so that $\boldsymbol{\mu}'_+ = \boldsymbol{S}$, a nonsingular diagonal matrix. Applying Theorem 4.1 to the matrices $\boldsymbol{\eta}_k = \boldsymbol{\mu}'_k\boldsymbol{S}^{-1}$ gives Corollary 4.3 below:

COROLLARY 4.3. *Let $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ be the singular value decomposition of $\boldsymbol{\mu}_+$. Define*

$$\underset{F\times F}{\boldsymbol{\eta}_k} = \underset{F\times I}{\boldsymbol{U}^T}\,\underset{I\times J}{\boldsymbol{\mu}_k}\,\underset{J\times F}{\boldsymbol{V}}\,\underset{F\times F}{\boldsymbol{S}^{-1}}, \qquad k = 1, \ldots, K.$$

Assume that the trilinear matrix equations (4) *and the parameterization conventions* PC1–PC3 *hold.*

1. *If identifiability conditions* IC1 *and* IC2 *are satisfied, then the columns of $\boldsymbol{Z} := \boldsymbol{U}^T\boldsymbol{A}$ are (right) eigenvectors of $\boldsymbol{\eta}_k$ with eigenvalues equal to the diagonal elements of $\boldsymbol{D}_k\boldsymbol{D}_+^{-1} = \boldsymbol{\Lambda}_k$ for every $k = 1, \ldots, K$.*

2. *If all three identifiability conditions* IC1–IC3 *hold, then the columns of $\boldsymbol{Z}$ are the only common eigenvectors of $\boldsymbol{\eta}_k$, $k = 1, \ldots, K$.*

Theorem 4.1 or Corollary 4.3 can be used to demonstrate that the trilinear matrix equations (4) do not hold. The matrices $\boldsymbol{\theta}_k, k = 1, \ldots, K$, need not be symmetric, so that they can fail to have real eigenvalues. Even if the K matrices have real eigenvalues, the eigenvectors can fail to be common.

In some settings the matrices $\boldsymbol{\mu}_k$ are square and symmetric. When all K matrices are symmetric, we shall use the suggestive notation $\boldsymbol{\Sigma}_k$ for these matrices. (The application suggested by this notation is beyond the scope of this paper.) If the trilinear matrix equations (4) hold, then $\boldsymbol{\Sigma}_k = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{A}^T$, so that $\boldsymbol{A} = \boldsymbol{B}$ and $\boldsymbol{\Sigma}_+ = \boldsymbol{A}\boldsymbol{D}_+\boldsymbol{A}^T$. If we let the singular-value decomposition of $\boldsymbol{\Sigma}_+$ be given by $\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T$, the symmetric matrices

$$\boldsymbol{\Psi}_k = \boldsymbol{S}^{-1/2}\boldsymbol{U}^T\boldsymbol{\Sigma}_k\boldsymbol{U}\boldsymbol{S}^{-1/2}, \qquad k = 1, \ldots, K,$$

have the property that $\boldsymbol{\Psi}_+ = \boldsymbol{I}_F$. Corollary 4.4 below states that the decomposition can be obtained from the $\boldsymbol{\Psi}_k$'s, so that the eigenvalues needed are those of symmetric matrices.

COROLLARY 4.4. *If the K matrices $\boldsymbol{\Sigma}_k$ are square symmetric matrices and if* IC1 *and* IC3 *hold, then the only common eigenvectors of $\boldsymbol{\Psi}_k$ are proportional to the columns of $\boldsymbol{X}$, where*

$$\boldsymbol{X} = \boldsymbol{S}^{-1/2}\boldsymbol{U}^T\boldsymbol{A}.$$

The eigenvalues are the diagonal elements of the matrices $\boldsymbol{D}_k\boldsymbol{D}_+^{-1}$.

Proof. The columns of $\boldsymbol{X}$ will not generally be of unit length, so that normalizations will be required in order to meet the parameterization conventions PC1–PC3. The demonstration that the columns of $\boldsymbol{X}$ are eigenvectors uses the two representations for $\boldsymbol{\Sigma}_+^+$, one formula in terms of $\boldsymbol{A}^+$ and one formula in terms of $\boldsymbol{U}$. It is necessary to show that $\boldsymbol{\Psi}_k\boldsymbol{X} = \boldsymbol{X}\boldsymbol{\Lambda}_k$. The first step is to substitute for $\boldsymbol{\Sigma}_k$ in the left-hand side:

$$\begin{aligned}\boldsymbol{\Psi}_k\boldsymbol{X} &= (\boldsymbol{S}^{-1/2}\boldsymbol{U}^T(\boldsymbol{A}\boldsymbol{D}_k\boldsymbol{A}^T)\boldsymbol{U}\boldsymbol{S}^{-1/2})\boldsymbol{S}^{-1/2}\boldsymbol{U}^T\boldsymbol{A} \\ &= \boldsymbol{S}^{-1/2}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{D}_k\boldsymbol{A}^T\boldsymbol{U}\boldsymbol{S}^{-1}\boldsymbol{U}^T\boldsymbol{A}.\end{aligned} \tag{12}$$

The product of the first three matrices is just $\boldsymbol{X}$. By Proposition 2.1, $\boldsymbol{U}\boldsymbol{S}^{-1}\boldsymbol{U}^T = (\boldsymbol{\Sigma}_+)^+$. Since $(\boldsymbol{\Sigma}_+)^+ = (\boldsymbol{A}\boldsymbol{D}_+\boldsymbol{A}^T)^+$, Proposition 2.1 also implies that $(\boldsymbol{\Sigma}_+)^+$ is $(\boldsymbol{A}^T)^+\boldsymbol{D}_+^{-1}\boldsymbol{A}^+$. Therefore, the matrix products to the right of $\boldsymbol{D}_k$ in (12) are $\boldsymbol{A}^T\boldsymbol{U}\boldsymbol{S}^{-1}\boldsymbol{U}^T\boldsymbol{A} = \boldsymbol{A}^T(\boldsymbol{A}^T)^+\boldsymbol{D}_+^{-1}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{D}_+^{-1}$. Therefore, $\boldsymbol{\Psi}_k\boldsymbol{X} = \boldsymbol{X}\boldsymbol{D}_k\boldsymbol{D}_+^{-1}$, as claimed.

The remainder of the proof is similar to the proof of the second part of Theorem 4.1, and so the details will be omitted. □

5. A decomposition algorithm.

In this section we present an algorithm that applies the decomposition. In the first subsection we list steps that will determine the parameters of an F-factor PARAFAC model in the absence of noise. We then indicate several modifications that will not change the results obtained when $\boldsymbol{\mu}$ satisfies the trilinear model. As described in §5.3 below, we have observed that a few of these modifications result in substantial improvements when applied to real or simulated arrays of fluorescence data. Subsection 5.2 discusses our implementation of this algorithm.

5.1. Steps in the algorithm.

1. Sum over the third index of the data array $\boldsymbol{Y}$ to obtain $\boldsymbol{Y}_+ = \sum_{k=1}^{K} \boldsymbol{Y}[,,k]$.
2. Fix $F > 1$.

When the array to be decomposed satisfies the trilinear model (3) and if PC3 holds, the rank of $\boldsymbol{\mu}_+$ will be F. If the context of the model ensures that all elements of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{\Gamma}$ are nonnegative, then PC3 will hold for the smallest F such that the trilinear model (3) holds. If $\boldsymbol{Y}$, the array to be decomposed, is a perturbation of an array $\boldsymbol{\mu}$ satisfying (3), the rank of $\boldsymbol{Y}_+$ need not be F. The value of F should be at least as large as the number of dominant singular values of $\boldsymbol{Y}_+$.

3. Obtain the F leading singular values and vectors of $\boldsymbol{Y}_+$. Let $\boldsymbol{S}$ be the F-by-F diagonal matrix whose diagonal elements are the singular values, let $\boldsymbol{U}$ be the I-by-F matrix whose orthonormal columns are the left singular vectors, and let $\boldsymbol{V}$ be the J-by-F matrix whose orthonormal columns are the right singular vectors.
4. Form $\hat{\boldsymbol{\eta}}_k = \boldsymbol{U}^T\boldsymbol{Y}[,,k]\boldsymbol{V}\boldsymbol{S}^{-1}$, $k = 1, \ldots, K$.
5. Find a K-vector of weights $\boldsymbol{w}$ such that $\hat{\boldsymbol{\eta}}(\boldsymbol{w}) = \sum_{k=1}^{K} w[k]\hat{\boldsymbol{\eta}}_k$ has F distinct real eigenvalues.
6. Let $\boldsymbol{Z}(\boldsymbol{w})$ be an F-by-F matrix whose columns are unit-length eigenvectors of $\hat{\boldsymbol{\eta}}(w)$.
7. Take $\hat{\boldsymbol{A}} = \boldsymbol{U}\boldsymbol{Z}(\boldsymbol{w})$.
8. Find

$$\boldsymbol{B}_o^T = \hat{\boldsymbol{A}}^+\boldsymbol{Y}_+ = (\boldsymbol{Z}(\boldsymbol{w}))^{-1}\boldsymbol{S}\boldsymbol{V}^T.$$

   The second equality follows from application of Proposition 2.1 to step 7.

9. Set $\boldsymbol{D}_+$ to be the diagonal matrix of lengths of columns of $\boldsymbol{B}_o$, and take $\hat{\boldsymbol{B}} = \boldsymbol{B}_o\boldsymbol{D}_+^{-1}$.
10. Find $\boldsymbol{D}_k = (\boldsymbol{Z}(\boldsymbol{w}))^{-1}\hat{\boldsymbol{\eta}}_k\boldsymbol{Z}(\boldsymbol{w})\boldsymbol{D}_+$ and set the kth row of $\hat{\boldsymbol{\Gamma}}$ equal to the diagonal of $\boldsymbol{D}_k$, $k = 1, \ldots, K$.
11. Permute the columns of $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$, and $\hat{\boldsymbol{\Gamma}}$, and change the signs of the columns to satisfy the parameterization conventions PC2 and PC3.

When $F = 1$, the matrices $\hat{\boldsymbol{\eta}}_k$, $k = 1, \ldots, K$, formed in step 4 are scalars, so that all linear combinations $\hat{\boldsymbol{\eta}}(\boldsymbol{w})$ automatically have a single ($F = 1$) nontrivial real eigenvalue. For the $F = 1$ PARAFAC model, steps 4–11 can be replaced by step 4_1:

4_1. Set $\hat{\boldsymbol{A}} = \boldsymbol{U}$, $\hat{\boldsymbol{B}} = \boldsymbol{V}$, and $\hat{\Gamma}[k, 1] = \hat{\boldsymbol{A}}^T\boldsymbol{Y}[,,k]\hat{\boldsymbol{B}}$, $k = 1, \ldots, K$.

This modification is an application of (9) with $\boldsymbol{\mu}_k$ replaced by $\boldsymbol{Y}_k$ because the Moore–Penrose generalized inverse of an n-by-1 unit vector is the 1-by-n row vector that is its transpose.

The matrices $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ are characterized differently in the steps above, even though the trilinear model (3) and the parameterization conventions PC1–PC3 treat the first two ways symmetrically. The proposition below establishes that this asymmetry is only in appearances: Exchanging the first two ways of an array will always give equivalent decompositions even if $\boldsymbol{Y}$ does not satisfy the trilinear model. Thus steps 8 and 9 are equivalent to 8′ and 9′ provided that the vector $\boldsymbol{w}$ of step 8′ is the vector used in step 5.

8′. Form $\hat{\boldsymbol{\eta}}^* = \boldsymbol{V}^T(\boldsymbol{Y}[,,k])^T\boldsymbol{U}\boldsymbol{S}^{-1}$, and set $\hat{\boldsymbol{\eta}}^*(\boldsymbol{w}) = \sum_{k=1}^{K} w[k]\hat{\boldsymbol{\eta}}_k^*$.

9′. Let $\boldsymbol{Z}^*$ be an F-by-F matrix whose columns are unit-length (right) eigenvectors of $\hat{\boldsymbol{\eta}}^*$. Take $\hat{\boldsymbol{B}} = \boldsymbol{V}\boldsymbol{Z}^*$.

PROPOSITION 5.1. *Let $\boldsymbol{Y}$ be an I-by-J-by-K data array. Define a J-by-I-by-K array $\boldsymbol{Y}^*$ by setting $\boldsymbol{Y}^*[,,k] := (\boldsymbol{Y}[,,k])^T$. Assume it is possible to take $\boldsymbol{w}$ such that step* 5 *is possible when applied to $\boldsymbol{Y}$. Let $\hat{\boldsymbol{B}}$ be the matrix obtained in step* 9 *when the algorithm is applied to $\boldsymbol{Y}$, and let $\hat{\boldsymbol{A}}^*$ be the matrix of parameters for the nominal first way when the algorithm is applied to $\boldsymbol{Y}^*$ (with the same vector $\boldsymbol{w}$). Then $\hat{\boldsymbol{B}} = \hat{\boldsymbol{A}}^*$.*

Proof. The superscript * will denote quantities derived from the permuted array $\boldsymbol{Y}^*$. To establish the proposition, the results of each $\boldsymbol{Y}^*$ step must be written in terms of quantities computed from $\boldsymbol{Y}$. Since $\hat{\boldsymbol{A}}^* = \boldsymbol{U}^*\boldsymbol{Z}^*$ in step 7, the matrices $\boldsymbol{U}^*$ and $\boldsymbol{Z}^*$ need to be written in terms of unstarred quantities. The K-vector $\boldsymbol{w}$ will be omitted from the notation since the same $\boldsymbol{w}$ is used with both $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$.

In step 1, $\boldsymbol{Y}^*_+ = \boldsymbol{Y}^T_+$, so that the diagonal matrices of singular values in step 3 are equal and left and right singular vectors are exchanged: $\boldsymbol{S}^* = \boldsymbol{S}$, $\boldsymbol{U}^* = \boldsymbol{V}$, and $\boldsymbol{V}^* = \boldsymbol{U}$. In step 4, $\hat{\boldsymbol{\eta}}^*_k = \boldsymbol{V}^T(\boldsymbol{Y}_k)^T\boldsymbol{U}\boldsymbol{S}^{-1} = \boldsymbol{S}(\boldsymbol{S}^{-1}\boldsymbol{V}^T\boldsymbol{Y}^T_k\boldsymbol{U})\boldsymbol{S}^{-1} = \boldsymbol{S}\hat{\boldsymbol{\eta}}^T_k\boldsymbol{S}^{-1}$. In step 5, $\hat{\boldsymbol{\eta}}^* = \boldsymbol{S}(\hat{\boldsymbol{\eta}}^T)\boldsymbol{S}^{-1}$ or the matrix $\hat{\boldsymbol{\eta}}^*$ is similar to $\hat{\boldsymbol{\eta}}^T$ with diagonal similarity matrix $\boldsymbol{S}^{-1}$ [12, p. 44]. Therefore, $\hat{\boldsymbol{\eta}}^*$ and $\hat{\boldsymbol{\eta}}$ have the same eigenvalues, so that $\hat{\boldsymbol{\eta}}^*$ has F distinct positive real eigenvalues.

The matrix equation summarizing the fact that the columns of $\boldsymbol{Z}$ are (right) eigenvectors of $\hat{\boldsymbol{\eta}}$ with diagonal matrix of eigenvalues $\boldsymbol{\Lambda}$ is $\hat{\boldsymbol{\eta}}\boldsymbol{Z} = \boldsymbol{Z}\boldsymbol{\Lambda}$. When the eigenvalues are distinct, then $\boldsymbol{Z}$ is nonsingular, so that $\hat{\boldsymbol{\eta}} = \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^{-1}$ and $\hat{\boldsymbol{\eta}}^T = (\boldsymbol{Z}^T)^{-1}\boldsymbol{\Lambda}\boldsymbol{Z}^T$, implying that $\hat{\boldsymbol{\eta}}^T(\boldsymbol{Z}^T)^{-1} = (\boldsymbol{Z}^T)^{-1}\boldsymbol{\Lambda}$ or that the columns of $(\boldsymbol{Z}^T)^{-1}$ are eigenvectors of $\hat{\boldsymbol{\eta}}^T$. But the similarity of $\hat{\boldsymbol{\eta}}^*$ and $\hat{\boldsymbol{\eta}}^T$ implies that the columns of $\boldsymbol{S}(\boldsymbol{Z}^T)^{-1}$ are (unnormalized) eigenvectors of $\boldsymbol{\eta}^*$, so that the normalized eigenvectors $\boldsymbol{Z}^*$ are $\boldsymbol{S}(\boldsymbol{Z}^T)^{-1}\boldsymbol{D}^{-1}$, where $\boldsymbol{D}$ is the diagonal matrix of the lengths of the columns of $\boldsymbol{S}(\boldsymbol{Z}^T)^{-1}$. Since $\boldsymbol{V}$ is an orthogonal matrix, the lengths of the rows of $\boldsymbol{S}(\boldsymbol{Z}^T)^{-1}$ are the same as the lengths of the rows of $\boldsymbol{S}(\boldsymbol{Z}^T)^{-1}\boldsymbol{V}$, implying that $\boldsymbol{D} = \boldsymbol{D}_+$. Now substituting for $\boldsymbol{U}^*$ and $\boldsymbol{Z}^*$, we obtain $\boldsymbol{Z}^* = \boldsymbol{V}\boldsymbol{S}(\boldsymbol{Z}^T)^{-1}\boldsymbol{D}_+^{-1} = \boldsymbol{B}_o\boldsymbol{D}_+^{-1} = \hat{\boldsymbol{B}}$. The last equality is from step 9, and the preceding equality is from step 8. These equations complete the proof of the proposition. □

The computation of $\hat{\boldsymbol{\Gamma}}$ in step 10 relies on the fact that the columns of $\boldsymbol{Z}(\boldsymbol{w})$ will be eigenvectors of $\hat{\boldsymbol{\eta}}_k$ if the trilinear matrix equations (4) hold. But when $\boldsymbol{Y}$ does not satisfy the trilinear model (3), the matrices $\boldsymbol{D}_k$ may not be diagonal and the estimate of $\boldsymbol{\Gamma}$ may not be satisfactory. Since this decomposition is often intended to initialize a nonlinear least-squares calculation, step 10 can be replaced by step 10′ in the determination of the $\hat{\boldsymbol{\Gamma}}$ that gives the best least-squares fit to $\boldsymbol{Y}$. By the trilinear matrix equations (4), $\boldsymbol{\mu}[,,k]$ is a linear function of $\boldsymbol{\Gamma}[k,]$ when $\boldsymbol{A}$ and $\boldsymbol{B}$ are known, so that a conditional least-squares step will give $\boldsymbol{\Gamma}$. If the trilinear matrix equations (4) hold, steps 10 and 10′ are equivalent, but the numerical results below demonstrate that 10′ should be preferred to 10 for initializing nonlinear least-squares fit.

10′. Estimate $\boldsymbol{\Gamma}[k,]$ from $\boldsymbol{Y}[,,k]$ by using least-squares regression of the model (4) with $\boldsymbol{A} = \hat{\boldsymbol{A}}$ and $\boldsymbol{B} = \hat{\boldsymbol{B}}$.

This variant illustrates the nonuniqueness of decomposition algorithms: Either step 10 or step 10′ will recover $\boldsymbol{\Gamma}$ if (4) holds, but the algorithms differ away from the model.

5.2. Our implementation of the algorithm. We have implemented the decomposition in FORTRAN by using IMSL subroutines [13]. We use the decomposition to initialize an alternating least-squares algorithm equivalent to the iterative steps described by Carroll and Chang [4] and by Harshman [10]. Sands and Young [25] give equivalent algorithms involving slightly more compact expressions by exploiting the multilinear structure. Our implementation does not use the flexibility offered by our formulation of step 5 because we have generally found that one or more $\hat{\boldsymbol{\eta}}_k$ has F real eigenvalues. We restrict our attention to weight vectors $\boldsymbol{w}$ containing one 1 and $K-1$ 0's, so that $\hat{\boldsymbol{\eta}}(\boldsymbol{w})$ is always one of the matrices $\hat{\boldsymbol{\eta}}_k$. The simplest choice of weights is to select the k with largest minimum distance between F real eigenvalues. In §5.3.1 below the choice of k is delayed until after one cycle of an alternating least-squares algorithm has been completed.

The alternating least-squares algorithms exploit the conditional linearity of multilinear models. The alternating least-squares algorithm has been extremely slow for fluorescence data sets, and we now use ad hoc acceleration methods. At each step these methods extrapolate the recent parameter estimates to approximate the parameters attaining the minimum. Iteration stops when the ratio of the vector norm of the difference between the current estimate and the extrapolated estimate to the vector norm of the current parameter estimates is less than 10^{-4}.

When $F = 1$ the initialization is not used. We have found that the alternating least-squares algorithm converges rapidly (always in five or fewer cycles) and that the limit points and the speed of convergence are not especially sensitive to the initialization.

We have run this code on an IBM 3081D computer at the Instructional and Research Computing Center, Ohio State University; on Pyramid 90X, DEC3100, and DEC5400 computers in the Mathematics and Statistics Computer Laboratory at Ohio State University; and on CRAY computers (XMP; later YMP) at the Ohio Supercomputer Center. Like most programs in active use, this code evolves. The code implementing the algorithm and the tests described in §5.3 below is in [18, Appendix 1].

5.3. Our experience. We now review our experience with this decomposition to initialize iterative nonlinear least-squares calculations. Our experience with five real data sets is described in §5.3.1. Four sets of fluorescence data were collected from aqueous solutions of biochemicals. The rationale for applying the trilinear model to fluorescence data is explained in [17]. Our fifth data set is the tongue-position data reported by Harshman, Ladefoged, and Goldstein [11]. Fluorescence data were also simulated for a study reported in [16]. The simulation results are described in §5.3.2. Except for the tongue data set, the three ways correspond to excitation wavelength, emission wavelength, and concentration of a chemical that affects fluorescence intensity. We focus on the behavior of the decomposition and its variants as an initialization of nonlinear least-squares calculations.

5.3.1. Performance of decomposition on real data. Each of the real fluorescence data sets has some structurally missing entries because fluorescence intensity cannot be measured accurately when excitation and emission wavelengths are very close. Since complete arrays are required to attempt the decomposition and since the fraction of entries missing is small, we filled the missing entries with the expected fluorescence intensity determined from the parameter estimates from the model with $F-1$ when initializing the computations for the model with F terms. The constructed entries are used *only* for the decomposition; all iterative least-squares steps use only the observed data.

Table 1 is a summary of the behavior of the algorithm for five real data sets. The triple of integers below the code name of each data set is the dimension vector of the three-way data array. The negative number below the vector indicates the number of missing elements. For more description of the fluorescence data sets, see [18, Appendix 2].

We attempted K decompositions for each data set and for $F = 2, 3, 4$. For each decomposition the matrix $\hat{\eta}(w)$ of step 5 was set equal to one of the matrices $\hat{\eta}_k$ of step 4. All eigenvalues from step 5 are listed in the tables of [18, Appendix 2]. When any eigenvalues were complex, step 5 failed. The number of slabs for which step 5 failed is reported in the column of Table 1 headed "Failure." The number of failures of step 5 increases with F, the number of terms fit. None of the five decompositions

TABLE 1

Summary of real data sets. Four sets of fluorescence data from aqueous solutions of biochemicals were collected in the laboratory of the second author. The fifth data set is the tongue-position data of [11]. *The triple of integers below the code name of each data set is the dimension vector of the three-way data array. The negative number below the vector indicates the number of missing elements. For explanation of the four sums-of-squares* SS1–SS4, *see text.*

Data set	F	Failure	SS1	Slab	SS2	Slab	SS3	Slab	SS4
PEA	1								2.91
(17,15,9)	2	0	4.15	(3)	.96	(1)	.44	(1)	.42
−(8 × 9)	3	3	4.83	(3)	.87	(6)	.31	(9)	.27
	4	5	58.86	(2)	.58	(9)	.25	(2)	.17
PCGK	1								3.237
(11,22,7)	2	1	.573	(6)	.568	(3)	.503	(2)	.498
−(17 × 7)	3	4	.510	(3)	.442	(2)	.429	(5)	.422
	4	4	.669	(5)	.394	(5)	.354	(5)	.306
NAT	1								197.23
(8,22,7)	2	0	15.78	(3)	14.42	(3)	5.96	(3)	5.41
−(8 × 7)	3	0	.81	(2)	.61	(2)	.50	(2)	.47
	4	2	1.12	(3)	.47	(2)	.33	(3)	.19
LADH	1								15.84
(15,27,7)	2	0	8.79	(1)	8.67	(1)	6.15	(1)	5.36
−(16 × 7)	3	0	2.47	(5)	1.13	(3)	.65	(1)	.38
	4	2	1.69	(1)	.49	(1)	.40	(1)	.34
TONG	1								240.56
(10,13,5)	2	1	20.94	(2)	20.62	(4)	20.16	(4)	19.99
−(0)	3	2	24.43	(3)	18.88	(1)	17.82	(4)	18.43
	4	5							

attempted provided an initialization for the tongue data set (TONG) with $F = 4$. We note that in the original source the model with $F = 2$ was selected.

For every k for which step 5 was possible, the remaining steps were performed. The smallest residual sum of squares from the parameter estimates at the end of step 10 is designated SS1 in Table 1, with the number of the corresponding slab indicated in parentheses. The smallest residual sum of squares resulting when step 10′ replaces step 10 is designated SS2. Note that the slab corresponding to SS1 is not always the slab corresponding to the smallest SS2. The smallest residual sum of squares after one ALS cycle beyond the step giving SS2 is designated SS3. The alternating least-squares algorithm was iterated to convergence for the slab corresponding to SS3. The residual sum of squares at convergence is designated SS4. For the residual sums of squares for every slab, see [18, Appendix 2].

5.3.2. Performance of decomposition on simulated fluorescence data. We extracted some results from a small simulation study reported in [16], a preliminary study of several methods of determining F in the trilinear model (3). We first outline the simulation, and we then report the results and compare them with the real data.

For the simulation, five independent data arrays were generated for each of three deterministic 10-by-12-by-5 arrays $\boldsymbol{\mu}$ satisfying the trilinear model. Each deterministic array corresponds to plausible fluorescence parameters. The number of terms in the true array $\boldsymbol{\mu}$ was $\tilde{F} = 1, 2, 3$. The three theoretical arrays are referred to as NFAK1, NFAK2, and NFAK3, with the number denoting the number of terms in the true array $\boldsymbol{\mu}$. The elements of $\boldsymbol{\mu}$ were between 0 and 10^4. The data arrays were simulated by adding independent pseudorandom normal variables with mean 0 and standard deviation 100. For all 15 simulated data arrays the trilinear model was fit with F set

TABLE 2

*Summary of simulated data sets. Slab number is the value of k used in step 4. The entry in each cell of the table is the number (out of 5) of replicated data arrays for which step 5 was possible. The * denotes models and slabs for which the computer code gave complex eigenvalues even when no noise was present.*

		Slab number				
Model	F	1	2	3	4	5
NFAK1	2	4	4	5	5	5
	3	2*	1*	3*	4	2
	4	1*	1	1*	3*	1*
NFAK2	2	5	5	3	5	5
	3	5	4	5	5	5
	4	2*	2	3	3*	3
NFAK3	2	5	5	5	5	5
	3	5	5	3	5	5
	4	5	2	4	5	5

equal to 1, 2, 3, and 4.

Table 2 gives the number of slabs that generated successful step 5's. Figure 1 shows the actual eigenvalues: Each column of plots corresponds to a different theoretical array; each row of plots corresponds to a different number of terms fitted. The circles in each panel are sample eigenvalues (when real) plotted against slab number. The eigenvalues of the theoretical arrays are superimposed crosses. It is evident from the figure that the sample eigenvalues were close to the true eigenvalues when the correct number of terms were fit and when separation of the true eigenvalues was larger than the variation in the sample eigenvalues. When more eigenvalues were fit than were present, some of the sample eigenvalues were close to the population parameters but the other sample eigenvalues varied considerably. Examination of the three sums-of-squares fits showed that using conditional least-squares fits for $\hat{\boldsymbol{\Gamma}}$ is important.

6. Extensions to four or more ways. In this section we prove that arrays with four ways can be decomposed by using the decomposition for three-way arrays. We state an extension to more ways.

The four-way extension of the trilinear model (3) to an I-by-J-by-L-by-M four-way array $\boldsymbol{\mu}$ is called the *quadrilinear model*:

$$\boldsymbol{\mu} = \sum_{f=1}^{F} \boldsymbol{\alpha}_f \times \boldsymbol{\beta}_f \times \boldsymbol{\gamma}_f \times \boldsymbol{\delta}_f, \tag{13}$$

where the $\boldsymbol{\alpha}_f$'s, the $\boldsymbol{\beta}_f$'s, the $\boldsymbol{\gamma}_f$'s, and the $\boldsymbol{\delta}_f$'s are the F columns of the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Gamma}$, and $\boldsymbol{\Delta}$, respectively. The matrix $\boldsymbol{A}$ has I rows, $\boldsymbol{B}$ has J rows, $\boldsymbol{\Gamma}$ has L rows, and $\boldsymbol{\Delta}$ has M rows. The parameterization convention PC1 can be extended to include the third way by requiring that the columns of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{\Gamma}$ all have unit length. The sign convention of PC2 can be extended to the matrix $\boldsymbol{B}$, so that $\boldsymbol{\Gamma}$ carries the sign.

The quadrilinear model (13) can be written in terms of three-way arrays. For example, if the levels of the last two ways of the four-way array define levels of a way with $K = LM$ levels, a three-way I-by-J-by-K array $\tilde{\boldsymbol{\mu}}$ is induced by

$$\tilde{\mu}[i,j,k] = \mu[i,j,l,m], \qquad k = l + L(m-1), \quad l = 1,\ldots,L; \quad m = 1,\ldots,M.$$

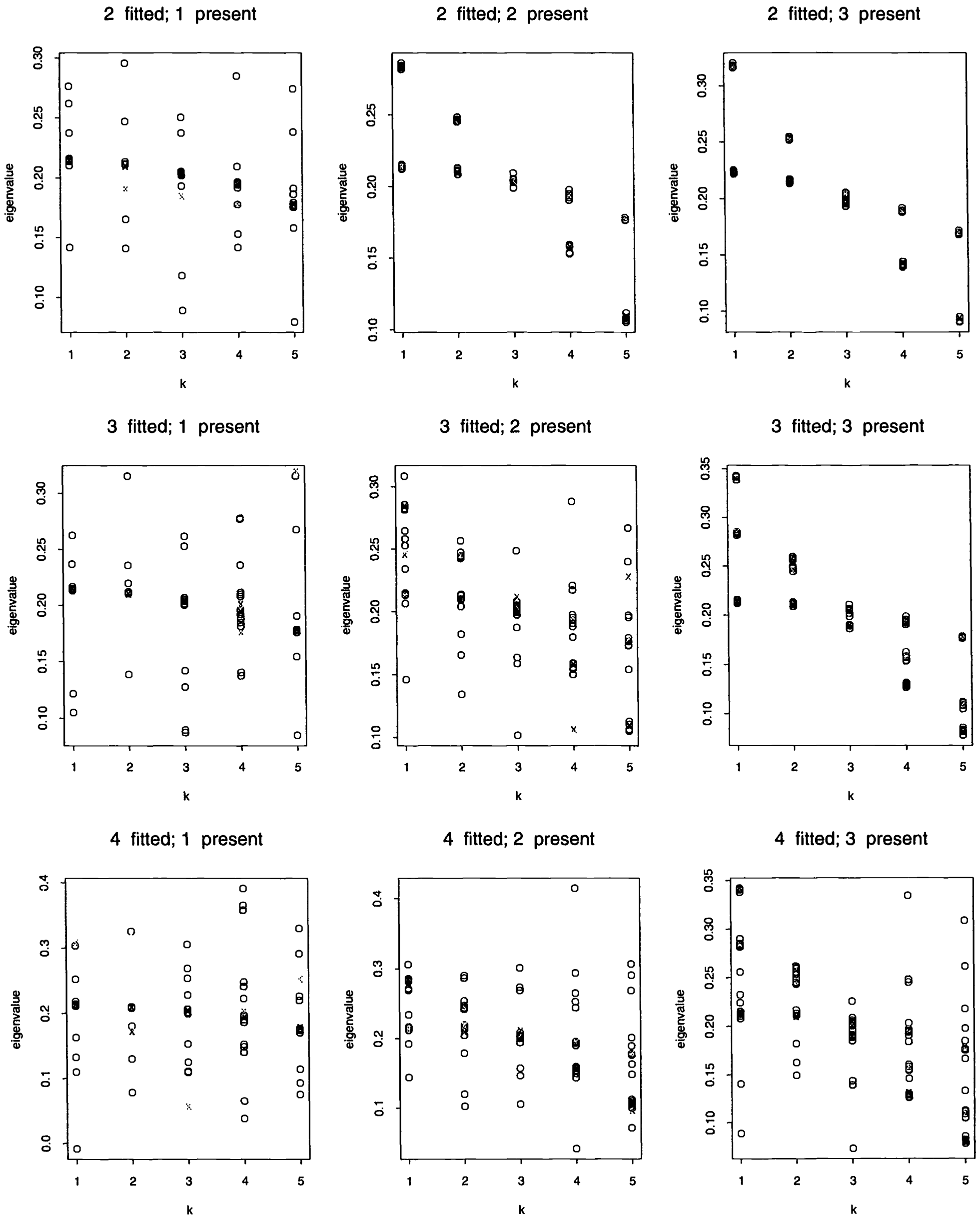


FIG. 1. *Eigenvalues from simulated arrays. Crosses are eigenvalues from theoretical arrays; open circles are eigenvalues from data arrays.*

The quadrilinear model (13) implies that the three-way array $\tilde{\boldsymbol{\mu}}$ satisfies

$$\tilde{\boldsymbol{\mu}} = \sum_{f=1}^{F} \boldsymbol{\alpha}_f \times \boldsymbol{\beta}_f \times \tilde{\boldsymbol{\gamma}}_f, \tag{14}$$

where $\tilde{\boldsymbol{\gamma}}_f$ is the K-vector equal to the Kronecker product of the L-vector $\boldsymbol{\gamma}_f$ with the M-vector $\boldsymbol{\delta}_f$, denoted $\boldsymbol{\gamma}_f \otimes \boldsymbol{\delta}_f$ and taken to have the first vector associated with the larger pattern and the second vector with the subpattern. Equation (14) has the form of the trilinear model (3), so that $\tilde{\boldsymbol{\mu}}$ has an F-factor three-way PARAFAC decomposition.

Identifiability conditions IC1 and IC2 are unaffected by the reduction from four ways to three ways. Identifiability condition IC3 applies directly to the K-vectors $\tilde{\boldsymbol{\gamma}}_1 \cdots \tilde{\boldsymbol{\gamma}}_F$. Lemma 6.1 at the end of this section shows that IC3 for $\tilde{\boldsymbol{\Gamma}}$ is equivalent to IC$_4$3 and IC$_4$4 below, so that the following *four-way identifiability conditions* are sufficient for identifiability when the parameterization conventions PC1–PC3 alluded to above are made:

(IC$_4$1) The columns of $\boldsymbol{A}$ are linearly independent.
(IC$_4$2) The columns of $\boldsymbol{B}$ are linearly independent.
(IC$_4$3) No two columns of $\boldsymbol{\Gamma}$ are collinear.
(IC$_4$4) No two columns of $\boldsymbol{\Delta}$ are collinear.

To determine the L-by-F matrix $\boldsymbol{\Gamma}$ and the M-by-F matrix $\boldsymbol{\Delta}$ from the K-by-F matrix $\tilde{\boldsymbol{\Gamma}}$, the rows of $\tilde{\boldsymbol{\Gamma}}$ are summed to add up the coefficients from $\boldsymbol{\Delta}$ (or from $\boldsymbol{\Gamma}$), leaving the coefficients from $\boldsymbol{\Gamma}$ (respectively, $\boldsymbol{\Delta}$). Explicitly, form the L-by-F matrix $\boldsymbol{G}$ by summing consecutive blocks of M rows of $\tilde{\boldsymbol{\Gamma}}$:

$$G[l, f] = \sum_{m=1}^{M} \tilde{\Gamma}[(l-1)M + m, f], \qquad l = 1, \ldots, L, \quad f = 1, \ldots, F.$$

The quadrilinear model (13) then implies that

$$\boldsymbol{G}[, f] = \boldsymbol{\Gamma}[, f] \mathbf{1}_M^T \boldsymbol{\Delta}[, f], \qquad f = 1, \ldots, F. \tag{15}$$

Since the columns of $\boldsymbol{\Gamma}$ are assumed to have length 1, $\boldsymbol{\Gamma}$ can now be determined from (15). Once the matrix $\boldsymbol{\Gamma}$ is known, the matrix $\boldsymbol{\Delta}$ can be found by summing every Mth row of $\tilde{\boldsymbol{\Gamma}}$ to get the M-by-B matrix $\tilde{\boldsymbol{G}}$:

$$\tilde{G}[m, f] = \sum_{l=1}^{L} \tilde{\Gamma}[(l-1)M + m, f], \qquad m = 1, \ldots, M, \quad f = 1, \ldots, F.$$

It is easy to verify that $\boldsymbol{\Delta}[, f] = \tilde{\boldsymbol{G}}[, f] / \mathbf{1}_L^T \boldsymbol{\Gamma}[, f], f = 1, \ldots, F$.

We now sketch the extension of the identifiability conditions to N-way arrays, where $N \geq 3$. Let $\boldsymbol{\mu}$ be an N-way array with ν_n levels in the nth way such that a formula like (13) holds. Assume that the N matrices $\boldsymbol{\theta}_n$ with ν_n rows and F columns containing the parameters for the nth way satisfy parameterization conventions like the parameterization conventions PC1–PC3, that is, the first $N - 1$ matrices will be assumed to have length-1 columns and a sign convention will be assumed. Induction on N shows that if none of the matrices $\boldsymbol{\theta}_n$, $n = 1, \ldots, N$, have any collinear columns and if at least two of the matrices have linearly independent columns, then the N matrices $\boldsymbol{\theta}_n$ can be identified from $\boldsymbol{\mu}$.

We conclude this section with a proof of the lemma containing the mathematical details.

Lemma 6.1. *If $\tilde{\gamma}_f = \gamma_f \otimes \delta_f$, $f = 1, 2$, and if the lengths of $\tilde{\gamma}_1$ and of $\tilde{\gamma}_2$ are positive, then $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are collinear K-vectors if and only if γ_1 and γ_2 are collinear L-vectors and δ_1 and δ_2 are collinear M-vectors.*

Proof. Suppose that $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are collinear. Then

$$\tilde{\gamma}_1 = c\tilde{\gamma}_2$$

for some $c \neq 0$. Since $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are Kronecker products, the L M-dimensional subvectors of γ_f are just

$$\gamma_1[l]\delta_1 = c\gamma_2[l]\delta_2, \qquad l = 1, \ldots, L.$$

Observe that $\gamma_1[l] = 0$ if and only if $\gamma_2[l] = 0$. Since the length of the Kronecker product of two vectors is the product of their lengths, there must be at least one l such that $\gamma_1[l] \neq 0$. For all such l

$$\delta_1 = c\frac{\gamma_2[l]}{\gamma_1[l]}\delta_2.$$

That is, the vector δ_1 is a multiple of δ_2, or δ_1 and δ_2 are collinear. The collinearity of γ_1 and γ_2 follows from a similar argument. □

7. Discussion.

The essential features of the decomposition have been in both the psychometric and chemometric literatures concerning initialization of alternating least-squares algorithms. All of these papers discuss real data sets as well. We briefly compare their approaches with ours, and we then summarize our conclusions and recommendations. We conclude our discussion with a brief comparison of array and tensor decompositions.

The psychometric literature discusses several types of initialization, all of which are surveyed and compared for a setting with symmetric slabs by Carroll, DeSoete, and Pruzansky [6]. One simple method is to use one or more random starts. Harshman [10] recommends several random starts. Carroll, DeSoete, and Prozansky found this method to be dominated by other methods, and so did we. In our applications, the inner products of the columns of $\boldsymbol{A}$ (and of $\boldsymbol{B}$) often exceed 0.9; such configurations are not likely to arise from simple random schemes. Another simple method is to attempt to use bilinear models, such as the singular vectors of $\boldsymbol{Y}_{+}$. We found that imposing nonnegativity constraints did not provide useful initializations. A third approach is to postulate that the identifiability conditions almost fail or that the columns of $\boldsymbol{A}$, $\boldsymbol{B}$, or $\boldsymbol{\Gamma}$ are very close to a prespecified lower-dimensional subspace. In this approximate setting, Carroll, Pruzansky, and Kruskal's [5] linearly constrained PARAFAC model (CANDELINC) applies to reduce to another unconstrained PARAFAC model for an array of smaller dimension. We have not applied this approach directly because in our applications submodels can be specified only for $\boldsymbol{\Gamma}$, the parameter matrix with the smallest number of rows. However, reduction to η_k in Corollary 4.3 can be thought of as postulating data-dependent linear models. We do not, however, pursue alternating least-squares fits for these models. Other initialization methods are more similar to ours. The proposal of Sands and Young [25] differs in detail because they find the eigenvalues of

$$\frac{1}{K}\sum_{k=1}^{K}\hat{\eta}_k^r,$$

where r is an integer (often 2) rather than finding the eigenvalues of a matrix $\hat{\eta}(\boldsymbol{w})$.

The chemometric literature has also used initializations based on eigenvalues. Sanchez and Kowalski [22] rediscovered the approach for I-by-J-by-2 arrays. Sanchez and Kowalski extend their earlier paper to I-by-J-by-K arrays in [21], [23], [24]. They also propose finding the eigenvectors of a matrix. To present their algorithm in an extension of our notation, define the IJ-by-K matrix $\tilde{\boldsymbol{\mu}}$ by stacking the matrices $\tilde{\boldsymbol{\mu}}_k, k = 1, \dots, K$, one above the next. Let $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ be the first and second right singular vectors of $\tilde{\boldsymbol{\mu}}$. Sanchez and Kowalski use the eigenvectors of $\tilde{\boldsymbol{\mu}}(\boldsymbol{w}_2)^+\tilde{\boldsymbol{\mu}}(\boldsymbol{w}_1)$ to recover the decomposition. When noise is present Sanchez and Kowalski first project $\boldsymbol{Y}$ onto a larger class of multilinear arrays, a class that includes the PARAFAC models. Their preprocessing is different from ours, so that their algorithm will not behave exactly like ours. Our experiments with their approach suggest that their methods and ours behave similarly, with neither method dominating. Our approach has the advantage that we always try K decompositions, so that the impact of complex eigenvalues may be avoided.

Burdick, Tu, McGown, and Millican [3] also describe the same eigendecomposition; readers who compare our paper with theirs should notice that our f (respectively, F) is their k (respectively, K) and our k (respectively, K) is their f (respectively, F). They do not apply the decomposition directly to $\boldsymbol{Y}$ but first apply alternating eigenproblems to obtain better estimates of the column spaces of $\boldsymbol{A}$ and $\boldsymbol{B}$ and then apply the eigendecomposition to $\tilde{\boldsymbol{Y}}_k = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{Y}_k\boldsymbol{V}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ have orthonormal columns and are such that their column spaces are estimates of the column spaces of $\boldsymbol{A}$ and $\boldsymbol{B}$. They are iteratively refining preprocessing as used by Sanchez and Kowalski [21]. Our preliminary experiments with our data indicate that the iterative eigendecomposition is more unstable than our direct initialization.

Sample variation from the trilinear model (3) can cause the decomposition to fail because complex eigenvalues can occur at step 5. Complex eigenvalues were observed for some k and some F for every data set. However, the imaginary part of the eigenvalues was not large: The absolute value of the ratio of the imaginary part to the real part of the complex eigenvalues never exceeded 0.189 for the fluorescence data sets. From simulations we see that complex eigenvalues can appear even when the correct model is fitted, so that complex eigenvalues should *not* be regarded as evidence that the trilinear model is inadequate. The set of arrays $\boldsymbol{Y}$ having real eigenvalues for every slab is a closed set, and those arrays having a slab generating at least one real eigenvalue with multiplicity at least 2 are on the boundary of this set, so that slight perturbations give arrays with complex roots. Indeed, slight perturbations of arrays having distinct real eigenvalues, some of which are not well separated, can also induce some complex roots. We have seen that nonunique eigenvalues occur when certain properties of the spectra nearly tie for two factors for some k. These ties can occur without the spectra being otherwise pathological, and so we recommend that several decompositions be attempted routinely. Table 1 shows that some decompositions are better than others when applied to real data, even though the decompositions give the same results when applied to arrays that exactly satisfy the trilinear model. We found the same comparison for simulated data. The original decomposition (with step 10) cannot be recommended because SS1 is smaller than SS4 from $F-1$ only for $F = 2$ (four data sets) and $F = 3$ (two data sets). The minimum SS1 fails to decrease as F increases for four of the five data sets. Sometimes SS1 is very large in real data sets: Two slabs of the PEA data exhibited this problem. Using step 10′ produces a substantial improvement: Only for $F = 3, 4$ of the PEA data does the smallest SS2 exceed SS4 from $F-1$. The conditional-least-squares property guarantees that SS3 will be less than SS2 (which is itself automatically less than SS1), and it is encouraging that the minimum SS3 was always less than SS4 from $F-1$, although the largest SS3 can exceed SS4 from $F-1$: This occurred in slab 5 of the PEA data when $F=4$.

A three-way array can be regarded as an array of components with respect to specified bases of a tensor in a tensor product of three finite-dimensional vector spaces [19]. If the bases in the vector spaces are unchanged, the array of components will change, although the underlying tensor does not. Decomposable tensors generate rank-1 arrays, and every tensor in a tensor product of finite-dimensional vector spaces can be written as a finite sum of decomposable tensors. If a tensor is written as a sum of decomposable tensors, a multilinear model is determined for each array of components of the tensor. However, determining the minimal number of terms in and the uniqueness of a decomposition is not routine in spaces that are tensor products of more than two vector spaces. For example, Kruskal, Harshman and Lundy introduced a 2-by-2-by-2 array (reproduced in [27]) that cannot be written as the sum of two rank-1 arrays. The array can be written as the sum of rank-1 arrays, but not uniquely, even under parameterization conventions PC1–PC3. This example contrasts with standard results for tensor products of two vector spaces (such as [19, p. 36]) and demonstrates again that decompositions for three-way arrays are not immediate from two-way results.

Acknowledgment. Particularly thorough comments from a referee have led us to improve the exposition and strengthen a lemma.

REFERENCES

[1] C. Appellof and E. Davidson, *Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents*, Anal. Chem., 53 (1981), pp. 2053–2056.

[2] P. Arabie, J. D. Carroll, and W. S. DeSarbo, *Three-Way Scaling and Clustering*, Quantitative Applications in the Social Sciences, Vol. 65, Sage, Newbury Park, CA, 1987.

[3] D. Burdick, X. Tu, L. McGown, and D. Millican, *Resolution of multicomponent fluorescent mixtures by analysis of the excitation-emission-frequency array*, J. Chemometrics, 4 (1990), pp. 15–28.

[4] J. D. Carroll and J. Chang, *Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart–Young" decompositions*, Psychometrika, 35 (1970), pp. 283–319.

[5] J. D. Carroll, S. Pruzansky, and J. B. Kruskal, CANDELINC: *A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters*, Psychometrika, 45 (1980), pp. 3–24.

[6] J. D. Carroll, G. DeSoete, and S. Pruzansky, *An evaluation of five algorithms for generating an initial configuration for* SINDSCAL, J. Classification, 6 (1989), pp. 105–119.

[7] R. Coppi and S. Bolasco, eds., *Multiway Data Analysis*, North-Holland, New York, 1989.

[8] M. Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press, Orlando, FL, 1984.

[9] T. Greville, *Note on the generalized inverse of a matrix product*, SIAM Rev., 8 (1966), pp. 518–521.

[10] R. Harshman, *Foundations of the* PARAFAC *procedure: Models and conditions for an "exploratory" multi-mode factor analysis*, UCLA Working Papers in Phonetics, 16 (1970), pp. 1–84.

[11] R. Harshman, P. Ladefoged, and L. Goldstein, *Factor analysis of tongue shapes*, J. Acoust. Soc. Am., 62 (1977), p. 693.

[12] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985; corrected reprint, 1990.

[13] IMSL, Inc., MATH/LIBRARY *User's Manual*, Houston, TX.

[14] J. B. KRUSKAL, *Three-way arrays: Rank and uniqueness of trilinear decompositions with application to arithmetic complexity and statistics*, Linear Algebra Appl., 18 (1977), pp. 95–138.

[15] H. LAW, C. SNYDER, JR., J. HATTIE, AND R. MCDONALD, EDS., *Research Methods for Multimode Data Analysis*, Praeger, New York, 1984.

[16] S. LEURGANS AND R. ROSS, *Cross-validation for multilinear models: Applications to biophysics*, 1989. Presented at Joint Statistical Meeting, Washington, DC.

[17] ———, *Multilinear models: Applications in spectroscopy*, Statist. Sci., 7 (1992), pp. 289–319.

[18] S. LEURGANS, R. ROSS, AND R. ABEL, *A Decomposition for* 3-*Way Arrays*, Tech. Report 448, Department of Statistics, Ohio State University, Columbus, OH, 1990.

[19] M. MARCUS, *Finite Dimensional Multilinear Algebra, Part* I, Marcel Dekker, New York, 1973.

[20] R. PRINGLE AND A. RAYNER, *Generalized Inverse Matrices with Applications to Statistics*, Hafner, New York, 1971.

[21] E. SANCHEZ AND B. KOWALSKI, *Tensorial resolution: A direct trilinear decomposition*, J. Chemometrics, 4 (1990), pp. 29–45.

[22] E. SANCHEZ AND B. R. KOWALSKI, *Generalized rank annihilation factor analysis*, Anal. Chem., 58 (1986), pp. 496–499.

[23] ———, *Tensorial calibration:* I. *First-order calculations*, J. Chemometrics, 2 (1988), pp. 247–263.

[24] ———, *Tensorial calibration:* II. *Second-order calculations*, J. Chemometrics, 2 (1988), pp. 265–290.

[25] R. SANDS AND F. W. YOUNG, *Component models for three-way data: An alternating least squares algorithm with optimal scaling features*, Psychometrika, 45 (1980), pp. 39–67.

[26] G. STYAN, *Generalized inverses*, in Encyclopedia of Statistical Sciences, Vol. 3, S. Kotz, N. Johnson, and C. Read, eds., John Wiley & Sons, Inc., New York, 1983, pp. 334–337.

[27] J. M. TEN BERGE, H. A. KIERS, AND J. DE LEEUW, *Explicit* CANDECOMP/PARAFAC *solutions for a contrived* $2 \times 2 \times 2$ *array of rank three*, Psychometrika, 53 (1988), pp. 579–589.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1084–1095, October 1993

012

SOME RESULTS ON THE NUMERICAL RANGE OF A DERIVATION*

NATÁLIA BEBIANO†, CHI-KWONG LI‡, AND JOÃO DA PROVIDÊNCIA§

Abstract. Let $A \in \mathbb{C}_{n\times n}$. For $1 \le p \le q \le n$, the (p,q) numerical range of A is defined and denoted by

$$W_{p,q}(A) = \{E_p(X^*AX) : X \in \mathbb{C}_{n\times q}, X^*X = I_q\},$$

where $E_p(Y)$ denotes the pth elementary symmetric function on the eigenvalues of Y. In this paper, the authors explore the relation of this concept and quantum physics, then prove some theorems relating the geometrical properties of the set $W_{p,q}(A)$ and the algebraic properties of the matrix A. In particular, it is shown that if $W_{p,q}(A)$ contains some special boundary points, then A is unitarily similar to a direct sum of matrices of lower dimensions. In some cases, the matrix A can be shown to be normal. A conjecture of Marcus on this subject is briefly discussed.

Key words. numerical range, derivation, compound matrix, Grassman space

AMS subject classification. 15A60

1. Introduction.

Let $A \in \mathbb{C}_{n\times n}$. For $1 \le p \le q \le n$ the (p,q) *numerical range* of A is defined and denoted by

$$W_{p,q}(A) = \{E_p(X^*AX) : X \in \mathbb{C}_{n\times q}, X^*X = I_q\},$$

where $E_p(Y)$ denotes the pth elementary symmetric function on the eigenvalues of Y. Let $Q_{q,n}$ denote the set of all strictly monotonic sequence of length q whose terms belong to the set $\{1,\ldots,n\}$. For $\omega \in Q_{q,n}$ and $Y \in \mathbb{C}_{n\times n}$, let $Y[\omega]$ be the $q \times q$ principal submatrix of Y lying in the rows and columns indexed by $\omega(1),\ldots,\omega(q)$. One easily verifies that

$$W_{p,q}(A) = \{E_p(UAU^*[\omega]) : \omega \in Q_{q,n}, UU^* = I_n\}.$$

Since $E_p(Y) = \operatorname{tr} C_p(Y)$, where $C_p(Y)$ denotes the pth compound of Y (e.g., see [7] for the definition and properties of compound matrices), one may replace $E_p(\cdot)$ by $\operatorname{tr} C_p(\cdot)$ in the above two definitions. The concept can be considered in $\wedge^q(\mathbb{C}^n)$, the qth Grassmann space over $\mathbb{C}^n$. In fact, if $D_p^q(A)$ is the pth derivation of A in $\wedge^q(\mathbb{C}^n)$ determined by the formula

$$C_q(I + tA) = \sum_{r=0}^{q} t^r D_r^q(A),$$

*Received by the editors November 21, 1991; accepted for publication (in revised form) January 27, 1992.

†Departamento de Matemática, Universidade de Coimbra, 3000 Coimbra, Portugal.

‡Department of Mathematics, The College of William and Mary, Williamsburg, Virginia 23185 (ckli@cs.wm.edu). The work of this author was supported by National Science Foundation grant DMS-91-00344. This research was done while visiting the Department of Mathematics, the University of Coimbra in the Summer, 1991, and was supported by the University of Coimbra and by a grant from Junta Nacional de Investigação Científica e Tecnológica, Portugal.

§Departamento de Física, Universidade de Coimbra, 3000 Coimbra, Portugal.

then (e.g., see [12])

$$W_{p,q}(A) = \{\langle D_p^q(A)x^\wedge, x^\wedge\rangle : x^\wedge \text{ is a decomposable unit vector in } \wedge^q(\mathbb{C}^n)\}.$$

It is not hard to see that

$$\begin{aligned} W_{p,q}(A) &= \{\operatorname{tr}(D_p^q(A)C_q(XX^*)) : X \in \mathbb{C}_{n\times q}, X^*X = I_q\} \\ &= \{\operatorname{tr}(D_p^q(A)C_q(UJ_qU^*)) : UU^* = I_n\}, \end{aligned}$$

where $J_q = I_q \oplus 0_{n-q}$. The set $W_{p,q}(A)$ reduces to

(a) the *classical numerical range* $W(A)$ if $1 = p = q$,

(b) the set $\{E_p(A)\}$ if $q = n$,

(c) the *q-numerical range* $W_q(A)$ if $1 = p \le q$,

(d) the classical numerical range of $D_p^q(A)$ if $p \le q = n-1$,

(e) the qth *decomposable numerical range* $W_q^\wedge(A)$ if $p = q$.

All these special cases are quite well studied. In particular, many interesting results relating the geometrical properties of $W_{p,q}(A)$ and the algebraic properties of the matrix A have been obtained for those special cases (e.g., see [2], [4], [5], and their references). In this paper, we study similar problems for the other cases, namely, when $1 < p < q < n-1$. Before proving our results, we first explore the relation between the (p,q) numerical range of A and quantum physics in §2. Then we study some special boundary points of $W_{p,q}(A)$ in §3. Section 4 deals with those matrices A for which $W_{p,q}(A)$ is a singleton or a line segment. Further generalizations of the concept motivated by quantum physics are discussed in §5. A conjecture of Marcus on the subject is also studied.

Suppose A has eigenvalues $\lambda_1, \ldots, \lambda_n$. We shall denote by $H_{p,q}(A)$ the set of all numbers of the form $E_p(\lambda_{\omega(1)}, \ldots, \lambda_{\omega(q)})$ with $\omega \in Q_{q,n}$. (Note that here $E_p(\cdot)$ denotes the pth elementary symmetric function of the numbers.) It is known (e.g., see [12]) that $H_{p,q}(A)$ is the spectrum of $D_p^q(A)$, and $H_{p,q}(A) \subset W_{p,q}(A)$. If A is normal, then $W_{p,q}(A) \subset \operatorname{conv} H_{p,q}(A) = W(D_p^q(A))$. We shall reserve the symbol ι to denote the sequence $(1, \ldots, q)$ in $Q_{q,n}$ in our discussion.

2. Motivations from physics. Suppose $x \in \mathbb{C}^n$ is a unit vector representing a (normalized) state of a certain one-particle system, and the behavior of the particle is determined by the observable quantity represented by $A \in \mathbb{C}_{n\times n}$. Then $\langle Ax, x\rangle$ is the average value (or expectation value, in a probabilistic sense) of a measurement of the observable A in the state x, and the classical numerical range $W(A)$ can be regarded as the collection of all possible average values (expectation values) of measurements of A in states of the one-particle system under consideration. Since $W(A)$ is convex, if z_1 and z_2 are possible average values of measurements of A, then $\lambda z_1 + (1-\lambda)z_2$ for any $\lambda \in [0,1]$ is also a possible average value which can be attained by adjusting the state vector of the particle.

Suppose there are q particles in the system with (normalized) state vectors x_j, $1 \le j \le q$, such that $\{x_1, \ldots, x_q\}$ is an orthonormal set. Then their joint state is represented by $x^\wedge = x_1 \wedge \cdots \wedge x_q \in \wedge^q(\mathbb{C}^n)$. Observable quantities of an additive nature, such as the kinetic energy, linear momentum or angular momentum, occur frequently in physics. If $A \in \mathbb{C}_{n\times n}$ represents such a quantity for a one-particle system, then the corresponding quantity for a q-particle system is represented by $D_1^q(A)$. The average value of measurements of such a quantity in the state $x^\wedge$ is computed by

$$\langle D_1^q(A)x^\wedge, x^\wedge\rangle = \sum_{i=1}^q \langle x_1 \wedge \cdots \wedge Ax_i \wedge \cdots \wedge x_q, x^\wedge\rangle = \sum_{i=1}^q \langle Ax_i, x_i\rangle.$$

Thus $W_q(A)$ can be regarded as the collection of all possible average values of measurements of $D_1^q(A)$ for the system. Since $W_q(A)$ is always convex, the physical interpretation in the preceding paragraph also applies to this system.

Quantities of an interactive nature, such as a 2-body, 3-body or many-body interaction or operator, also occur in physics. A p-particle interaction is only effective in a q-particle system if $q \geq p$ and, in some simple circumstances, may be described as a derivation $D_p^q(A)$. (Under more complex circumstances it may be necessary to consider different operators $A_1, \dots, A_k$ acting simultaneously on different particles (see also §4), or linear combinations of derivations.) The average value of measurements of $D_p^q(A)$ in the state $x^\wedge$ is then computed by

$$\begin{aligned}\langle D_p^q(A)x^\wedge, x^\wedge\rangle &= \sum_{\omega\in Q_{p,q}} \langle x_1 \wedge \cdots \wedge Ax_{\omega(1)} \wedge \cdots \wedge Ax_{\omega(p)} \wedge \cdots \wedge x_q, x^\wedge\rangle \\ &= \sum_{\omega\in Q_{p,q}} \sum_{\sigma\in S_p} \prod_{i=1}^{p} \langle Ax_{\omega(\sigma(i))}, x_{\omega(i)}\rangle \operatorname{sgn}(\sigma),\end{aligned}$$

where S_p is the symmetric group of degree p and $\operatorname{sgn}(\sigma)$ denotes the sign of the permutation σ. Thus $W_{p,q}(A)$ can be regarded as the collection of all possible average values of measurements of $D_p^q(A)$ in the state $x^\wedge$ of such a system and is very useful in the study of it.

To better describe the physical interpretation of the operator $D_p^q(A)$, we introduce the concepts of *creation operator* $f_i : \wedge^q(\mathbb{C}^n) \to \wedge^{q+1}(\mathbb{C}^n)$ defined by $f_i(x^\wedge) = e_i \wedge x^\wedge$, where $\{e_1, \dots, e_n\}$ is the standard basis of $\mathbb{C}^n$, and *destruction operator* (or *annihilation operator*) $g_i : \wedge^{q+1}(\mathbb{C}^n) \to \wedge^q(\mathbb{C}^n)$, which is just the adjoint operator of f_i. One easily verifies the following properties of the creation and destruction operators:

(a) $f_i f_j + f_j f_i = 0$ for all i, j,

(b) $g_i g_j + g_j g_i = 0$ for all i, j,

(c) $f_i g_j + g_j f_i = \delta_{ij}$ for all i, j.

Creation and destruction operators are the building blocks of linear operators acting on a q-body system. In fact, every linear operator acting on a q-body system can be expressed as a linear combination of linear operators of the form $f_{i_1} \cdots f_{i_r} g_{j_s} \cdots g_{j_1}$. Such an operator destroys particles in states $j_1, \dots, j_s$ and creates particles in states $i_1, \dots, i_r$. In general (not always) the number of particles is conserved. Thus the number of particles destroyed and the number of particles created are the same, i.e., $r = s$. In our case, $D_1^q(A)$ is given by the formula $\sum a_{ij} f_i g_j$, and $D_p^q(A)$ is given by the formula

$$(p!)^{-1} \sum_{i_1, j_1, \dots, i_p, j_p} a_{i_1 j_1} \cdots a_{i_p j_p} f_{i_1} \cdots f_{i_p} g_{j_p} \cdots g_{j_1}.$$

One may see [3] for more information concerning creation and destruction operators.

3. Some special boundary points.

THEOREM 3.1. *Let $1 < p < q < n$ and let A be in upper triangular form. Suppose $\omega \in Q_{q,n}$ satisfies $E_p(A[\omega]) \in \partial W_{p,q}(A)$ and $D_{p-1}^{q-1}(A[\omega])$ is nonsingular. Then for any $i < j$ with $|\{i,j\} \cap \{\omega(1), \dots, \omega(q)\}| = 1$ we have $a_{ij} = 0$. In particular, if $\omega = \iota$, then $A = A[\iota] \oplus A_2$.*

Proof. Let ω satisfies the hypotheses of the theorem and suppose $z = E_p(A[\omega])$. We first show that $a_{ij} = 0$ if $j = i + 1$ and $|\{i,j\} \cap \{\omega(1), \dots, \omega(q)\}| = 1$. Suppose there is such an (i, j) pair with $a_{ij} \neq 0$. Then by the elliptical range theorem (e.g., see [14]),

$$W\left(\begin{bmatrix} a_{ii} & a_{ij} \\ 0 & a_{jj} \end{bmatrix}\right)$$

is a nondegenerate elliptical disk with a_{ii} and a_{jj} as foci. Thus there exists $r > 0$ such that whenever $\varepsilon \in \mathbb{C}$ with $|\varepsilon| < r$, we can find orthogonal vectors u, v in the linear span of $\{e_i, e_j\}$ satisfying $u^*Au = a_{ii} + \varepsilon$ and $v^*Av = a_{jj} - \varepsilon$. Let U be the unitary matrix obtained from the identity matrix by replacing the ith and jth columns by u and v. Let $\omega' \in Q_{q-1,n}$ be obtained from ω by removing k from its range, where $\{k\} = \{i,j\} \cap \{\omega(1), \ldots, \omega(q)\}$, and let $z' = E_{p-1}(A[\omega'])$. Then $E_p(U^*AU[\omega]) = z \pm \varepsilon z'$ depending on $k = i$ or $k = j$. Since $D_{p-1}^{q-1}(A[\omega])$ is nonsingular, $z' \neq 0$. It follows that for all ε with $|\varepsilon| < r$, $z + \varepsilon z' \in W_{p,q}(A)$, which contradicts the fact that $z \in \partial W_{p,q}(A)$.

Next consider those (i,j) pairs with $j = i+2$ and $|\{i,j\} \cap \{\omega(1), \ldots, \omega(q)\}| = 1$. Suppose there is such an (i,j) pair with $a_{ij} \neq 0$. Then by the same arguments as those in the preceding paragraph, we see that there exists $r > 0$ such that whenever $\varepsilon \in \mathbb{C}$ with $|\varepsilon| < r$, we can find orthogonal vectors u, v in the linear span of $\{e_i, e_j\}$ satisfying $u^*Au = a_{ii} + \varepsilon$ and $v^*Av = a_{jj} - \varepsilon$. Let U be the unitary matrix obtained from the identity matrix by replacing the ith and jth columns by u and v. Define $\omega' \in Q_{q-1,n}$ and $z' = E_{p-1}(A[\omega'])$ as in the preceding paragraph. Since we have shown that $a_{st} = 0$ if $t = s+1$ and $|\{s,t\} \cap \{\omega(1), \ldots, \omega(q)\}| = 1$, the matrix $U^*AU[\omega]$ is still in upper triangular form and $E_p(U^*AU[\omega]) = z \pm \varepsilon z'$. Since $D_{p-1}^{q-1}(A[\omega])$ is nonsingular, $z' \neq 0$. It follows that for all ε with $|\varepsilon| < r$, $z + \varepsilon z' \in W_{p,q}(A)$, which contradicts the fact that $z \in \partial W_{p,q}(A)$.

Repeating the above arguments for $j = i+3$ and so on, we get the conclusion.

COROLLARY 3.2. *Let $1 < p < q < n$. Suppose $A \in \mathbb{C}_{n\times n}$ is such that $D_{p-1}^{q-1}(A)$ is nonsingular. Set $k = \binom{n}{q} - 2\binom{n-2}{q-1}$.*

(a) *If A is in upper triangular form and there is more than k $\omega \in Q_{q,n}$ such that $E_p(A[\omega]) \in \partial W_{p,q}(A)$, then A is a diagonal matrix.*

(b) *If there are more than k eigenvalues of $D_p^q(A)$ (counting multiplicities) lying on $\partial W_{p,q}(A)$, then A is normal.*

Proof. (a) Suppose the hypotheses are satisfied. Then (see [10, Lemma 1]) for any $1 \leq i < j \leq n$ there is a $\beta \in Q_{q,n}$ such that $E_p(A[\beta]) \in \partial W_{p,q}(A)$ and the set $\{i,j\} \cap \{\beta(1), \ldots, \beta(q)\}$ is a singleton. Since $D_{p-1}^{q-1}(A)$ is nonsingular, so is $D_{p-1}^{q-1}(A[\beta])$. Thus the (i,j) entry of A equals zero. Hence A is a diagonal matrix.

(b) Let U be a unitary matrix such that $A' = U^*AU$ is in upper triangular form. Apply the result in (a) to A'. The result follows. □

Marcus and Sandy in [12] showed that for a normal matrix A, if $z = E_p(A[\iota])$ is a simple eigenvalue of $D_p^q(A)$ and z is a vertex of $\operatorname{conv} H_{p,q}(A) = W(D_p^q(A))$, then $A = A[\iota] \oplus A_2$. In the following result, we obtain the same conclusion without requiring A to be normal.

THEOREM 3.3. *Let $1 < p < q < n$ and let $A \in \mathbb{C}_{n\times n}$. Suppose $z = E_p(A[\omega])$ is a nondifferentiable boundary point of $W(D_p^q(A))$ for some $\omega \in Q_{q,n}$ and z is a simple eigenvalue of $D_p^q(A)$. Then for any $i \neq j$ with $|\{i,j\} \cap \{\omega(1), \ldots, \omega(q)\}| = 1$ we have $a_{ij} = 0$. In particular, if $\omega = \iota$, then $A = A[\iota] \oplus A_2$.*

Proof. For simplicity, we assume that $\omega = \iota$. Since $z = E_p(A[\iota])$ is a nondifferentiable boundary point of $W(D_p^q(A))$, $D_p^q(A) = [z] \oplus B$. Thus $e_1 \wedge \cdots \wedge e_q$ is a unit eigenvector of $D_p^q(A)$. Suppose U^*AU is in upper triangular form such that $E_p(U^*AU[\iota]) = z$, then $u_1 \wedge \cdots \wedge u_q$ is also a unit eigenvector of $D_p^q(A)$ corresponding to z, where u_i is the ith column of U. As a result, $e_1 \wedge \cdots \wedge e_q$ and $u_1 \wedge \cdots \wedge u_q$ are linear dependent, and thus $\{e_1, \ldots, e_q\}$ and $\{u_1, \ldots, u_q\}$ have the same linear span. Since $u_j^* A u_i = 0$ for all $1 \leq j \leq q < i \leq n$, we have $e_j^* A e_i = 0$ for all $1 \leq j \leq q < i \leq n$. Let $V = V_1 \oplus V_2$ be a unitary matrix such that $V_1 \in \mathbb{C}_{q\times q}$ and $A' = V^*AV$ is in upper triangular form. We have $z = E_p(A'[\iota]) \in \partial W_{p,q}(A)$. We claim that $D_{p-1}^{q-1}(A'[\iota])$ is nonsingular. If it is not true, we may assume that A' has diagonal entries $\lambda_1, \ldots, \lambda_n$ such that $E_{p-1}(\lambda_1, \ldots, \lambda_{q-1}) = 0$. It follows that $z = E_p(\lambda_1, \ldots, \lambda_q) = E_p(\lambda_1, \ldots, \lambda_{q-1}, \lambda_{q+1})$, which contradicts the assumption that z is a simple eigenvalue of $D_p^q(A)$. So our claim is true and we can apply Theorem 3.1 and conclude that $A' = A'[\iota] \oplus A'_2$. □

Let $S \subset \mathbb{C}$. Following Bebiano [1], we say that $z \in S$ is a corner of S if there exists a $r > 0$ such that the set $\{\nu \in S : |\nu - z| < r\}$ is contained in an angle smaller than π with z as the vertex. Note that if z is a corner of S and $z \in T \subset S$, then z is a corner of T.

THEOREM 3.4. *Let $1 < p < q < n$ and let $A \in \mathbb{C}_{n\times n}$. Suppose $E_p(A[\omega])$ is a corner of $W_{p,q}(A)$ for some $\omega \in Q_{q,n}$, and $D_{p-1}^{q-1}(A[\omega])$ is nonsingular. Then for any $i \neq j$ with $|\{i,j\} \cap \{\omega(1), \ldots, \omega(q)\}| = 1$ we have $a_{ij} = 0$. In particular, if $\omega = \iota$, then $A = A[\iota] \oplus A_2$.*

Proof. For simplicity, we assume that $\omega = \iota$. Suppose A is not of the form $A[\iota] \oplus A_2$. Then there exists an $(q+1) \times (q+1)$ principal submatrix A' of A containing $A[\iota]$ such that A' is not in the form of $A'[\iota] \oplus A'_2$. Let U be a $q \times q$ unitary matrix such that $U^*A[\iota]U$ is in lower triangular form. Set $V = U \oplus I_1$ and $B = V^*A'V$. Note that $E_p(B[\iota]) = E_p(A[\iota]) \in W_{p,q}(B) = W(D_p^q(B)) \subset W_{p,q}(A)$ is a corner of $W(D_p^q(B))$. Thus $E_p(B[\iota])$ is the only nonzero entry in the first row and first column of $D_p^q(B)$. Note that the $(1, q+1)$ entry of $D_p^q(B)$ is the coefficient of t^p in the polynomial

$$(-1)^{q+1} b_{1,q+1} t \prod_{i=2}^{q} (1 + b_{ii}t),$$

which equals $b_{1,q+1}E_{p-1}(b_{22}, \ldots, b_{qq})$. Since $E_{p-1}(b_{22}, \ldots, b_{qq})$ is an eigenvalue of $D_p^q(A[\iota])$ and is nonzero by assumption, it follows that $b_{1,q+1} = 0$.

Next, observe that the $(1, q)$ entry of $D_p^q(B)$ is the coefficient of t^p in the polynomial

$$(-1)^q b_{2,q+1} t(1 + b_{11}t) \prod_{i=3}^{q} (1 + b_{ii}t),$$

which equals $b_{1,q+1}E_{p-1}(b_{11}, b_{33}, \ldots, b_{qq})$. By the same arguments as before, we see that $b_{2,q+1} = 0$. Repeating the arguments, we see that $b_{i,q+1} = 0$ for all $i \leq q$.

Applying the same arguments to $b_{q+1,j}$ for $j = q, q-1, \ldots, 1$, we see that all of them equal zero. Thus $B = B[\iota] \oplus B_2$ and so must A'. This is a contradiction. □

COROLLARY 3.5. *Let $1 < p < q < n$ and let $A \in \mathbb{C}_{n\times n}$. Set $k = \binom{n}{q} - 2\binom{n-2}{q-1}$.*

(a) *Suppose there is more than k $\omega \in Q_{q,n}$ such that $D_{p-1}^{q-1}(A[\omega])$ is nonsingular and $E_p(A[\omega])$ is a corner of $\partial W_{p,q}(A)$, then A is a diagonal matrix.*

(b) *If $0 \notin W_{p-1,q-1}(A)$ and $W_{p,q}(A)$ has more than k corners, then A is normal.*

Proof. Similar to that of Corollary 3.2.

COROLLARY 3.6. *Let $1 < p < q < n$. Suppose $A \in \mathbb{C}_{n\times n}$ is such that $0 \notin W_{p-1,q-1}(A)$.*

(a) *The following conditions are equivalent:*

(i) $\operatorname{conv} W_{p,q}(A)$ *is a polygon (including interior),*

(ii) $\operatorname{conv} W_{p,q}(A) = \operatorname{conv} H_{p,q}(A)$,

(iii) $W_{p,q}(A) \subset \operatorname{conv} H_{p,q}(A)$.

(b) *The set $W_{p,q}(A)$ is a line segment if and only if A is normal and all points in $H_{p,q}(A)$ are collinear. In particular, $W_{p,q}(A)$ is a singleton if and only if A is a scalar matrix.*

Proof. (a) In general, $H_{p,q}(A) \subset W_{p,q}(A)$. Hence $\operatorname{conv} H_{p,q}(A) \subset \operatorname{conv} W_{p,q}(A)$. Since $0 \notin W_{p-1,q-1}(A)$, all the vertices of $\operatorname{conv} W_{p,q}(A)$ belong to $H_{p,q}(A)$ by Theorem 3.4. As a result, if condition (i) holds, then $\operatorname{conv} W_{p,q}(A) \subset \operatorname{conv} H_{p,q}(A)$ and hence condition (ii) holds. The converse is trivial. The equivalence of (ii) and (iii) is clear.

(b) If $W_{p,q}(A)$ is a line segment, then $H_{p,q}(A) \subset \partial W_{p,q}(A)$. By Corollary 3.2, A is normal. Clearly, the points in $H_{p,q}(A)$ must be collinear. Conversely, if A is normal and the points in $H_{p,q}(A)$ are collinear, then $W_{p,q}(A) = \operatorname{conv} H_{p,q}(A)$ is a line segment.

Finally, suppose $W_{p,q}(A)$ is a singleton. Then A is normal. If λ_1 and λ_2 are two distinct eigenvalues of A, then

$$E_p(\lambda_1, \lambda_3, \ldots, \lambda_{q+1}) - E_p(\lambda_2, \lambda_3, \ldots, \lambda_{q+1}) = (\lambda_1 - \lambda_2)E_{p-1}(\lambda_3, \ldots, \lambda_{q+1}) \neq 0$$

for $0 \notin W_{p-1,q-1}(A)$ by assumption. This is a contradiction. Thus A is a scalar matrix. The converse is clear.

Note that the assumptions such as $D_{p-1}^{q-1}(A)$ are nonsingular, $0 \notin W_{p-1,q-1}(A)$, in the theorems and corollaries are necessary. Otherwise, the conclusion may not hold as shown by the following example.

Example. Let $(p, q, n) = (2, 3, 4)$ and let $A = E_{13} + E_{24}$. Then $W_{p,q}(A) = \{0\} = H_{p,q}(A)$ and every point in $W_{p,q}(A)$ is a corner. However, A is not unitarily similar to $A_1 \oplus A_2$ with $A_1 \in \mathbb{C}_{3\times 3}$.

We shall study the matrices A for which $W_{p,q}(A)$ is a singleton or a line segment in §4. Note that the characterization of such matrices is not easy to obtain even with the assumption that $0 \notin W_{p-1,q-1}(A)$.

It is known that $|z| \leq E_p(\sigma_1(A), \ldots, \sigma_q(A))$ for any $z \in W_{p,q}(A)$, where $\sigma_1(A) \geq \cdots \geq \sigma_n(A)$ are the singular values of A. If there is $z \in W_{p,q}(A)$ such that $|z| = E_p(\sigma_1(A), \ldots, \sigma_q(A))$, then $z \in \partial W_{p,q}(A)$. For the characterization of those matrices A with $W_{p,q}(A)$ having such a boundary point, we have the following result (see [6, Thm. 8.1]).

THEOREM 3.7. *Let $1 < p < q < n$ and let $A \in \mathbb{C}_{n\times n}$ satisfy rank$(A) \geq q$. Suppose $z = E_p(A[\iota])$ satisfies $|z| = E_p(\sigma_1(A), \ldots, \sigma_q(A))$. Then $A = A[\iota] \oplus A_2$ such that $\mu A[\iota]$ is a positive semidefinite matrix with eigenvalues $\sigma_1(A), \ldots, \sigma_q(A)$.*

Note that if we do not assume that $\operatorname{rank}(A) \geq q$, the conclusion of the theorem may not hold. For example if $A = B \oplus 0_{n-p}$, where $B \in \mathbb{C}_{p\times p}$ is nonsingular, then $z = \det(B) = E_p(A[\iota]) = E_p(\sigma_1(A), \ldots, \sigma_q(A)) = \Pi_{i=1}^{p} \sigma_i(A)$. In fact, if $\operatorname{rank}(A) = r < q$, then $E_p(\sigma_1(A), \ldots, \sigma_q(A)) = E_p(\sigma_1(A), \ldots, \sigma_r(A))$. Moreover, if $z = E_p(X^*AX)$ satisfies $|z| = E_p(\sigma_1(A), \ldots, \sigma_q(A))$, where $X \in \mathbb{C}_{n\times q}$ satisfies $X^*X = I_q$, then X^*AX cannot have more than r nonzero eigenvalues and thus there exists $Y \in \mathbb{C}_{q\times r}$ with $Y^*Y = I_r$ such that $z = E_p(Y^*X^*AXY)$. Consequently, we have $z \in W_{p,r}(A)$ such that $|z| = E_p(\sigma_1(A), \ldots, \sigma_r(A))$, and Theorem 3.7 can be applied. (Note that in [6, Thm. 8.1, the condition "rank$(A) \geq q$" is incorrectly stated as "rank$(A) \geq p$".)

4. Degeneracy of $W_{p,q}(A)$.

The purpose of this section is to study the matrices A for which $W_{p,q}(A)$ is a singleton or a line segment. In [6, Thm. 7.1], it is shown that for $n > q \geq p \geq 1$, $H_{p,q}(A) = \{0\}$ if and only if A has less than p nonzero eigenvalues; for $n \geq p + q$, $W_{p,q}(A) = \{0\}$ if and only if $\operatorname{rank}(A) < p$; for $p + q > n$ and $q > p$, there exists a matrix A with $\operatorname{rank}(A) \geq p$ such that $W_{p,q}(A) = \{0\}$. We treat the case when $H_{p,q}(A) = \{\mu\}$ or $W_{p,q}(A) = \{\mu\}$ with $\mu \neq 0$ in the following theorem. Since the cases for $p = 1$ or $p = q$ are known, they are excluded in the following discussion.

THEOREM 4.1. *Let $1 < p < q < n$.*

(a) *Suppose $q < n - 1$. Then $H_{p,q}(A) = \{\mu\}$ with $\mu \neq 0$ if and only if all eigenvalues of A are equal. Consequently, $W_{p,q}(A) = \{\mu\}$ with $\mu \neq 0$ if and only if A is a scalar matrix.*

(b) *Suppose $q = n-1$ and $\mu \in \mathbb{C}$. There exists a nonscalar A such that $H_{p,q}(A) = W_{p,q}(A) = \{\mu\}$.*

Proof. (a) Suppose $n - 1 > q$ and suppose not all the eigenvalues are equal, say $\lambda_1 \neq \lambda_2$. Then for any $\omega \in Q_{q-1,n}$ with $\omega(1) \geq 3$, we have

$$\mu = E_p(\lambda_1, \lambda_{\omega(1)}, \dots \lambda_{\omega(q-1)}) = E_p(\lambda_2, \lambda_{\omega(1)}, \dots \lambda_{\omega(q-1)}).$$

It follows that $E_{p-1}(\lambda_{\omega(1)}, \dots \lambda_{\omega(q-1)}) = 0$. Since $n - 2 > q - 1$, we can apply [6, Thm. 7.1] to conclude that there are less than $p-1$ nonzero numbers among $\lambda_3, \dots, \lambda_n$. But then $E_p(\lambda_3, \dots, \lambda_{q+2}) = 0 \in H_{p,q}(A)$, which is a contradiction. The converse of the statement is clear.

Now suppose $W_{p,q}(A) = \{\mu\}$. Then all eigenvalues of A are equal. If A is not normal, then (e.g., see [11, Lemma 1]) A is unitarily similar to an upper triangular matrix A' whose $(1,2)$ entry is nonzero. Thus there exist orthonormal vectors u, v in the linear span of $\{e_1, e_2\}$ such that $v^*A'v$ is not an eigenvalue of A. Let U be a unitary matrix obtained from I by replacing its first two columns by u and v. Then $U^*A'U[2, \dots, q+1]$ is still in upper triangular form and $E_p(U^*A'U[2, \dots, q+1]) \neq \mu$, which is a contradiction. The converse of the statement is clear.

(b) Let $A = \nu(\frac{n-p-1}{1-p} I_2 \oplus I_{n-2})$ be such that $E_p(A[\iota]) = \mu$. Then for any unitary matrix U, the matrix $U^*AU[\iota]$ is unitarily similar to $\nu([\frac{n-p-1}{1-p}] \oplus [t] \oplus I_{n-3})$, where $(n-p-1)/(1-p) \leq t \leq 1$. Since $E_{p-1}(I_{n-3} \oplus [\frac{n-p-1}{1-p}]) = 0$,

$$\begin{aligned} E_p(U^*AU[\iota]) &= \nu^p \left\{ tE_{p-1}\left(I_{n-3} \oplus \left[\frac{n-p-1}{1-p}\right]\right) + E_p\left(I_{n-3} \oplus \left[\frac{n-p-1}{1-p}\right]\right) \right\} \\ &= \mu^p \left\{ E_p\left(I_{n-3} \oplus \left[\frac{n-p-1}{1-p}\right]\right) \right\} = \mu. \end{aligned}$$

The result follows.

Note that Theorem 4.1 improves the result in [6, Thm. 6.1]. Next we turn to those A such that $W_{p,q}(A)$ is a nondegenerate line segment.

THEOREM 4.2. *Suppose $1 < p < q$, $n \geq p + q$ and $A \in \mathbb{C}_{n\times n}$. Assume that $H_{p,q}(A)$ is not a singleton. Then all points in $H_{p,q}(A)$ are collinear if and only if one of the following conditions hold.*

(a) *A has exactly p nonzero eigenvalues.*

(b) *The eigenvalues of A lie on a straight line that passes through the origin.*

(c) *There are two distinct eigenvalues of A with multiplicities $n-1$ and 1, respectively.*

Consequently, $W_{p,q}(A)$ is a nondegenerate line segment if and only if one of the following holds.

(a′) *A is unitarily similar to $A_1 \oplus 0$, where $A_1 \in \mathbb{C}_{p\times p}$ is nonsingular.*

(b′) *A is normal and condition* (b) *or* (c) *holds.*

Proof. We prove the first statement by induction on p. First suppose $p = 2$. If A has three noncollinear eigenvalues, one of them must have multiplicity $n - 2$.

Otherwise, we can index the eigenvalues in such a way that $\lambda_1 \neq \lambda_2$ and $\lambda_3, \ldots, \lambda_n$ are not collinear. Since $E_2(\lambda_i, \lambda_{j_1}, \ldots, \lambda_{j_{q-1}})$ are collinear for all $3 \leq j_1 < \cdots < j_{q-1} \leq n$ and $i = 1, 2$, there exists a complex unit μ such that

$$\mu[E_2(\lambda_1, \lambda_{j_1}, \ldots, \lambda_{j_{q-1}}) - E_2(\lambda_2, \lambda_{j_1}, \ldots, \lambda_{j_{q-1}})] = \mu(\lambda_1 - \lambda_2)(\lambda_{j_1} + \cdots + \lambda_{j_{q-1}}) \in \mathbb{R}$$

for all $3 \leq j_1 < \cdots < j_{q-1} \leq n$. Hence $(\lambda_{j_1} + \cdots + \lambda_{j_{q-1}})$ are all collinear, which implies $\lambda_3, \ldots, \lambda_n$ are all collinear. This is a contradiction. Now suppose A has eigenvalues $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_3$. If $\lambda_3 = 0$, then condition (a) holds. Suppose $\lambda_3 \neq 0$. Since

$$z_1 = E_2(\lambda_1, \lambda_3, \ldots, \lambda_3), \quad z_2 = E_2(\lambda_2, \lambda_3, \ldots, \lambda_3), \quad z_3 = E_2(\lambda_3, \ldots, \lambda_3) \in H_{2,q}(A)$$

are collinear, there exists a complex unit μ such that

$$\mu(z_1 - z_3) = \mu(q-1)\lambda_3(\lambda_1 - \lambda_3), \ \mu(z_2 - z_3) = \mu(q-1)\lambda_3(\lambda_2 - \lambda_3) \in \mathbb{R}.$$

It follows that $\lambda_1, \lambda_2, \lambda_3$ are collinear, which is a contradiction.

Now suppose all the eigenvalues of A are collinear. If they lie on a line passing through the origin, then condition (b) holds. Suppose this is not the case. We claim that condition (c) holds. Assume that A has at least three distinct eigenvalues, say $\lambda_1, \lambda_2, \lambda_3$. One easily checks that for $1 \leq j < k \leq 3$ the points $\lambda_j \lambda_k$ are not collinear, and the points $(\lambda_j + \lambda_k) \sum_{s=4}^{q+1} \lambda_s$ are collinear. It follows that the points

$$E_2(\lambda_j, \lambda_k, \lambda_4, \ldots, \lambda_{q+1}) = \lambda_j \lambda_k + (\lambda_j + \lambda_k) \sum_{s=4}^{q+1} \lambda_s + E_2(\lambda_4, \ldots, \lambda_{q+1})$$

are not collinear. Thus A has only two distinct eigenvalues. If both eigenvalues have multiplicities larger than one, say $\lambda_1 = \lambda_2 \neq \lambda_3 = \lambda_4$, then one can check that the points

$$E_2(\lambda_1, \lambda_2, \lambda_5, \ldots, \lambda_{q+2}), \quad E_2(\lambda_2, \lambda_3, \lambda_5, \ldots, \lambda_{q+2}), \quad E_2(\lambda_3, \lambda_4, \lambda_5, \ldots, \lambda_{q+2}),$$

are not collinear. Thus one of the eigenvalues must have multiplicity one, and hence condition (c) holds.

Now suppose the statement holds for all (n', q', p') triples with $n' \geq p' + q'$, $q' > p' > 1$. Consider the statement for the (n, q, p) case with $p > p' \geq 2$.

If A has exactly p nonzero eigenvalues, then condition (a) holds. Suppose A has more than p nonzero eigenvalues. We claim that all eigenvalues of A are collinear. Suppose our claim is not true. We first show that one of the eigenvalues of A has multiplicity $n-2$. If it is not the case, then we can index the eigenvalues such that $\lambda_1 \neq \lambda_2$ and there are three noncollinear numbers among $\lambda_3, \ldots, \lambda_n$. Since $E_p(\lambda_i, \lambda_{j_1}, \ldots, \lambda_{j_{q-1}})$ are collinear for all $3 \leq j_1 < \cdots < j_{q-1} \leq n$ and $i = 1, 2$, there exists a complex unit μ such that $\mu(\lambda_1 - \lambda_2)E_{p-1}(\lambda_{j_1}, \ldots, \lambda_{j_{q-1}}) \in \mathbb{R}$ for all $3 \leq j_1 < \cdots < j_{q-1} \leq n$. Thus either all the points $E_{p-1}(\lambda_{j_1}, \ldots, \lambda_{j_{q-1}})$ are equal, or all of them lie on a line passing through the origin. By Theorem 4.1 or the induction assumption on the set $H_{p-1,q-1}(A')$, where $A' = \operatorname{diag}(\lambda_3, \ldots, \lambda_n)$, we conclude that $\lambda_3, \ldots, \lambda_n$ are collinear, which is a contradiction. Now suppose A has eigenvalues $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_3$. Since

$$z_1 = E_p(\lambda_1, \lambda_3, \ldots, \lambda_3), \quad z_2 = E_p(\lambda_2, \lambda_3, \ldots, \lambda_3), \quad z_3 = E_p(\lambda_3, \ldots, \lambda_3) \in H_{p,q}(A)$$

are collinear, there exists a complex unit μ such that

$$\mu(z_1 - z_3) = \mu\binom{q-1}{p-1}\lambda_3^{p-1}(\lambda_1 - \lambda_3),\ \mu(z_2 - z_3) = \mu\binom{q-1}{p-1}\lambda_3^{p-1}(\lambda_2 - \lambda_3) \in \mathbb{R}.$$

It follows that $\lambda_1, \lambda_2, \lambda_3$ are collinear, which is a contradiction. Thus our claim is proved.

Now suppose A has more than p nonzero eigenvalues and all of them are collinear. If they lie on a line passing through the origin, then condition (b) holds. Suppose this is not the case and assume $\lambda_1 \neq \lambda_2$. Then by arguments used previously, we see that either all the points $E_{p-1}(\lambda_{j_1}, \dots, \lambda_{j_{q-1}})$ are equal, or all of them lie on a line passing through the origin. Since there are at least $p-1$ nonzero numbers among $\lambda_3, \dots, \lambda_n$, by Theorem 4.1 or the induction assumption on the set $H_{p-1,q-1}(A')$, where $A' = \operatorname{diag}(\lambda_3, \dots, \lambda_n)$, we conclude that either (i) $\lambda_3 = \cdots = \lambda_n \neq 0$ or (ii) $\lambda_3 \neq \lambda_4 = \cdots = \lambda_n \neq 0$. Suppose (i) holds. If λ_1 or λ_2 equals λ_3, then condition (c) holds. Suppose $\lambda_1, \lambda_2, \lambda_3$ are distinct. One easily checks that for $1 \leq j < k \leq 3$ the points

$$E_p(\lambda_j, \lambda_k, \lambda_5, \dots, \lambda_{q+2}) = \lambda_j\lambda_k\binom{q-2}{p-2}\lambda_5^{p-2} + (\lambda_j + \lambda_k)\binom{q-2}{p-1}\lambda_5^{p-1} + \binom{q-2}{p}\lambda_5^p$$

are not collinear, which is a contradiction. Now suppose (ii) holds. If there are three distinct numbers among $\lambda_1, \dots, \lambda_4$, then we can show that the points in $H_{p,q}(A)$ are not collinear as in the previous case. Otherwise, we may assume $\lambda_1 = \lambda_3$ and $\lambda_2 = \lambda_4 = \cdots = \lambda_n$. But then one can check that the points

$$E_p(\lambda_1, \lambda_2, \lambda_5, \dots, \lambda_{q+2}), \quad E_2(\lambda_1, \lambda_3, \lambda_5, \dots, \lambda_{q+2}), \quad E_2(\lambda_2, \lambda_4, \lambda_5, \dots, \lambda_{q+2}),$$

are not collinear, which is a contradiction.

Now we turn to the set $W_{p,q}(A)$. Suppose $W_{p,q}(A)$ is a nondegenerate line segment. Since $H_{p,q}(A) \subset W_{p,q}(A)$, all points in $H_{p,q}(A)$ are collinear and they cannot be the same. As a result, one of the conditions (a) – (c) holds.

Suppose condition (a) holds. Then A is unitarily similar to $A' = \left[\begin{smallmatrix} A_1 & A_2 \\ 0 & A_3 \end{smallmatrix}\right]$ in upper triangular form such that $A_1 \in \mathbb{C}_{p\times p}$ with diagonal entries $\lambda_1, \dots, \lambda_p$ is nonsingular and A_3 is a nilpotent matrix. Suppose A_3 is nonzero. Then (see [11, Lemma 1]) we may assume that the $(1,2)$ entry of A_3 is nonzero. By the elliptical range theorem (e.g., see [13]), $W(A_3[1,2])$ is a nondegenerate circular disk centered at origin. Thus there exists $r > 0$ such that whenever $\varepsilon \in \mathbb{C}$ with $|\varepsilon| < r$, we can find orthogonal vectors u, v in the linear span of $\{e_{p+1}, e_{p+2}\}$ satisfying $u^*A'u = \varepsilon$ and $v^*A'v = -\varepsilon$. Let U be the unitary matrix obtained by replacing the $(p+1)$th and $(p+2)$th columns by u and v. Let $z = \det(A[2, \dots, q]) \neq 0$ and let $\omega = (2, \dots, p+1, p+3, \dots, q+2)$. Then $E_p(U^*A'U[\omega]) = \varepsilon z$. It follows that for all ε with $|\varepsilon| < r$, $\varepsilon z \in W_{p,q}(A)$, which contradicts the fact that $W_{p,q}(A)$ is a line segment. Now suppose A_2 is nonzero. We first show that the last row of A_2 is zero. Suppose there exists $p < j \leq n$ such that the (p,j) entry of A' is nonzero. Then by the same arguments as those in the preceding paragraph, we see that there exists $r > 0$ such that whenever $\varepsilon \in \mathbb{C}$ with $|\varepsilon| < r$, we can find orthogonal vectors u, v in the linear span of $\{e_p, e_j\}$ satisfying $u^*A'u = \lambda_p + \varepsilon$ and $v^*Av = -\varepsilon$. Let U be the unitary matrix obtained by replacing the pth and jth columns by u and v. Let $\omega \in Q_{q,n}$ be such that $\omega(i) = i$ for $i = 1, \dots, p$ and j does not belong to its range. Since we have shown that $A_3 = 0$, we see that $U^*A'U[\omega]$ is still in upper triangular form and $E_p(U^*A'U[\omega]) = \Pi_{i=1}^p\lambda_i + \varepsilon\Pi_{i=1}^{p-1}\lambda_i$. It follows that for all

ε with $|\varepsilon| < r$, $\Pi_{i=1}^{p}\lambda_i + \varepsilon\Pi_{i=1}^{p-1}\lambda_i \in W_{p,q}(A)$, which contradicts the fact that $W_{p,q}(A)$ is a line segment. Since we have shown that A_3 and the last row of A_2 are zero, we can apply the arguments to show that the $p-1$ row of A_2 is also zero. Repeating the arguments, we see that A_2 is zero. Hence condition (a') holds.

Now suppose condition (b) or (c) holds. If A is not normal, then (see [11, Lemma 1]) A is unitarily similar to A' in upper triangular form with diagonal entries $\lambda_1, \ldots, \lambda_n$ such that the $(1, 2)$ entry of A' is nonzero. Then by the same arguments as those in the preceding paragraph, we see that there exists $r > 0$ such that whenever $\varepsilon \in \mathbb{C}$ with $|\varepsilon| < r$, we can find orthogonal vectors u, v in the linear span of $\{e_1, e_2\}$ satisfying $u^*A'u = \lambda_1 + \varepsilon$ and $v^*A'v = \lambda_2 - \varepsilon$. Let U be the unitary matrix obtained by replacing the first and second columns by u and v. Since condition (b) or (c) holds, $H_{p-1,q-1}(A'[3, \ldots, n]) \neq \{0\}$. Thus there exists $\omega \in Q_{q,n}$ such that $\omega(1) = 2$ and $z' = E_{p-1}(A'[\omega(2), \ldots, \omega(q)]) \neq 0$. Moreover, $E_p(U^*A'U[\omega]) = E_p(A'[\omega]) - \varepsilon z'$. It follows that for all ε with $|\varepsilon| < r$, $E_p(A'[\omega]) - \varepsilon z' \in W_{p,q}(A)$, which contradicts the fact that $W_{p,q}(A)$ is a line segment. Thus A must be normal and hence condition (b') holds. □

Note that the conclusion in Theorem 4.2 does not hold if $q = n-1$. For example if A has two distinct eigenvalues, then $H_{p,q}(A)$ has collinear points. In fact, if $(p, q, n) = (2, 3, 4)$, then $H_{p,q}(A)$ has collinear points if A has eigenvalues (i) satisfying any one of the conditions (a) – (c) in Theorem 4.2; (ii) consisting of two distinct values; (iii) of the form $(\lambda_1, \lambda_1, \lambda_1\mu, \lambda_1\mu^{-1})$; (iv) $(\lambda_1, \lambda_1, -\lambda_1, \lambda_2)$; or (v) $(\lambda_1, -\lambda_1, \lambda_2, -\lambda_2)$. It would be interesting to characterize those matrices A for which $W_{p,q}(A)$ is a nondegenerate line segment without the assumption that $n \geq p + q$.

5. Further generalization and a conjecture of Marcus.

Recall that

$$W_{p,q}(A) = \{\operatorname{tr} C_p(J_qU^*AU) : U^*U = I\}.$$

Marcus [8] suggested the study of the set

$$W_p(A, B) = \{\operatorname{tr} C_p(BU^*AU) : U^*U = I\}.$$

Moreover, he conjectured that:

If A and B are normal, and if they are unitarily similar to the diagonal matrices D_A and D_B, respectively, then

$$W_p(A, B) \subset \operatorname{conv} H_p(A, B),$$

where

$$H_p(A, B) = \{\operatorname{tr} C_p(D_BP^tD_AP) : P \text{ a permutation matrix }\}.$$

It is known that the conjecture is valid if A or B equals J_q. Also, if the rank A or B is less than p, then $W_{p,q}(A) = H_{p,q}(A) = \{0\}$ and the conjecture is trivially satisfied. If rank of A or B is exactly p, then $W_{p,q}(A) \subset W(D_p^q(A)) = \operatorname{conv} H_{p,q}(A)$. Furthermore, since $W_{n-1}(A, B) = W_1(C_{n-1}(A), C_{n-1}(B)) \subset \operatorname{conv} H_{n-1}(A, B)$, the conjecture is valid for $p = n-1$. We have the following theorem.

THEOREM 5.1. *If A and B are positive semidefinite matrices with eigenvalues $\alpha_1 \geq \cdots \geq \alpha_n$ and $\beta_1 \geq \cdots \geq \beta_n$, respectively, then*

$$W_p(A, B) = \operatorname{conv} H_{p,q}(A) = [E_p(\alpha_1\beta_n, \ldots, \alpha_n\beta_1), E_p(\alpha_1\beta_1, \ldots, \alpha_n\beta_n)].$$

Proof. Let $f(x_1, \ldots, x_n) = E_p(x_1, \ldots, x_n)$. Apply Theorem 5 in [14], we get the result. □

Whether the conjecture is valid for hermitian matrices A and B is unknown. One possible approach for solving this special case is to show that the corners of $W_p(A,B)$ belong to $H_p(A,B)$. However, such a result is not available.

While $W_p(A,B)$ seems to be a nice generalization of $W_{p,q}(A)$, it does not have a natural physical interpretation. To give a more meaningful generalization of $W_{p,q}(A)$ in the context of quantum physics, consider

$$W_{p,q}(A) = \{\operatorname{tr}(D^q_p(A)C_q(U^*J_qU)) : U^*U = I\}.$$

One may regard $C_q(U^*J_qU)$ as the description of a pure state of a q-particle system in which every p of them interact with each other. A mixed state of the q particles is described by $C_q(U^*BU)$ for a fixed $B \in \mathbb{C}_{n\times n}$. In connection with such a state we can define

$$W^q_p(A,B) = \{\operatorname{tr}(D^q_p(A)C_q(U^*BU)) : U^*U = I\}.$$

Furthermore, one may consider a q-particle system with p-body interaction determined by the observable quantities represented by $A_1,\dots,A_k$. Then one would consider

$$W^q_\tau(A_1,\dots,A_k;B) = \{\operatorname{tr}(D^q_\tau(A_1,\dots,A_k)C_q(U^*BU)) : U^*U = I\},$$

where $\tau : p_1 + \cdots + p_k = q$ denotes a partition of q, and $D^q_\tau(A_1,\dots,A_k)$ is determined by the formula

$$C_q\left(\sum_{j=1}^{k} t_jA_j\right) = \sum_\tau t_1^{p_1}\cdots t_k^{p_k} D^q_\tau(A_1,\dots,A_k).$$

(See [9] for a particular case.) Clearly, $D^q_p(A) = D^q_\tau(A,I)$ with $\tau : p+(q-p)$. In terms of creation and destruction operators, $D^q_\tau(A_1,\dots,A_k)$ is determined by the formula

$$D^q_\tau(A_1,\dots,A_t) = (p_1!\cdots p_t!)^{-1} \sum_{i_1,j_1,\dots,i_p,j_p} b^{(1)}_{i_1j_1}\cdots b^{(p)}_{i_pj_p} f_{i_1}\cdots f_{i_p} g_{j_p}\cdots g_{j_1},$$

where $B_s = (b^{(s)}_{ij}) \in \mathbb{C}^{n\times n}$ with $1 \le s \le p$ such that

$$B_1 = \cdots = B_{p_1} = A_1, \qquad B_{p_1+1} = \cdots = B_{p_1+p_2} = A_2, \dots,$$

$$B_{p_1+\cdots+p_{t-1}+1} = \cdots = B_{p_1+\cdots+p_t} = A_k.$$

All these concepts seem to be interesting and deserve further study.

REFERENCES

[1] N. Bebiano, *Nondifferentiable points of $\partial W_c(A)$*, Linear and Mutlilinear Algebra, 19 (1986), pp. 249–257.

[2] N. Bebiano and C. K. Li, *A brief survey on decomposable numerical ranges*, Linear and Multilinear Algebra, 32 (1992), pp. 179–190.

[3] J. P. Blaizot and J. Ripka, *Quantum Theory of Finite Systems*, The MIT Press, Cambridge, MA, 1986.

[4] M. Goldberg, *On certain finite dimensional numerical ranges and numerical radii*, Linear and Multilinear Algebra, 7 (1979), pp. 329–342.

[5] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge Univ. Press, London, 1991.

[6] C. K. Li and N. K. Tsing, *The numerical range of derivations*, Linear Algebra Appl., 119 (1989), pp. 97–119.

[7] M. Marcus, *Finite Dimensional Multilinear Algebra*, Part I & II, Marcel Dekker, New York, 1973, 1975.

[8] ———, *Derivations, Plücker relations, and the numerical range*, Indiana Univ. Math. J., 22 (1973), pp. 1137–1149.

[9] ———, *Numerical ranges and higher derivations*, preprint.

[10] M. Marcus and I. Filippenko, *Nondifferentiable boundary points of the higher numerical range*, Linear Algebra Appl., 21 (1978), pp. 217–232.

[11] M. Marcus and M. Sandy, *Conditions for the generalized numerical range to be real*, Linear Algebra Appl., 71 (1985), pp. 219–239.

[12] ———, *Vertex points in the numerical range of a derivation*, Linear and Multilinear Algebra, 21 (1987), pp. 385–394.

[13] F. D. Murnaghan, *On the field of values of a square matrix*, Proc. Nat. Acad. Sci., 18 (1932), pp. 246–248.

[14] R. C. Thompson and S Therianos, *On the singular values of matrix products* – I, Scripta Mathematica, 29 (1971), pp. 99–110.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1096–1102, October 1993

FAST TRIANGULAR FACTORIZATION OF COVARIANCE MATRICES OF DIFFERENCED TIME SERIES*

RAJIV VIJAYAN[†] AND H. VINCENT POOR[‡]

Abstract. An algorithm is presented for computing the triangular factors of the covariance matrix of a random vector whose elements are the successive differences of a stationary time series. This algorithm has applications in the modeling of time series using a difference operator rather than using a shift operator, as is done in the conventional autoregressive model. The difference-operator based models offer the benefit of better numerical conditioning when applied to series that are obtained by sampling continuous-time processes at rapid rates. The covariance matrix of differenced data is not Toeplitz; however, it has a Toeplitz-like structure that is exploited in this paper to obtain an $O(n^2)$ algorithm. This algorithm is amenable to parallelization; i.e., it requires only $O(n)$ time when using n processors in parallel.

Key words. difference operators, triangular factorization, fast algorithms, Toeplitz-like matrices

AMS subject classifications. 65F30, 65Y05, 15-04, 15A23

1. Introduction. Recently, difference-operator based models have been proposed [9] as an alternative to conventional autoregressive models for modeling stationary time series that are obtained by sampling continuous-time processes at rates that are rapid relative to the process dynamics. In this approach, the model used is of the form

$$\delta^n X_k + \beta_{n,1}\delta^{n-1}X_k + \cdots + \beta_{n,n}X_k = \nu_k, \tag{1.1}$$

where $\{X_k\}$ is a zero-mean wide-sense stationary discrete-time random process obtained by sampling a continuous-time process $\{X(t)\}$ at interval Δ, and δ is an incremental difference operator defined by

$$\delta X_k = \frac{X_{k+1} - X_k}{\Delta} = \frac{(q-1)X_k}{\Delta}.$$

Here, q is the shift operator defined by $qX_k = X_{k+1}$, which forms the basis for describing the process dynamics in the conventional autoregressive model. From this definition of δ, we have

$$\begin{aligned}\delta^n X_k &= \frac{\delta^{n-1}X_{k+1} - \delta^{n-1}X_k}{\Delta}\\ &= \frac{1}{\Delta^n}\sum_{j=0}^{n}(-1)^j \binom{n}{j} X_{k+n-j}.\end{aligned}$$

The parameter vector $\underline{\beta}_n = [1, \beta_{n,1}, \ldots, \beta_{n,n}]^T$ that minimizes the mean-squared modeling error $E(\nu_k^2)$ in (1.1) satisfies

$$\mathbf{Q}_n\underline{\beta}_n = [\tilde{\pi}_n, 0, \ldots, 0]^T, \tag{1.2}$$

* Received by the editors March 11, 1991; accepted for publication (in revised form) February 4, 1992. This research was supported in part by U.S. Office of Naval Research grant N00014-89-J-1321 and in part by U.S. National Science Foundation grants NCR 90-02767 and CDA-91-21709.

[†] Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1101 W. Springfield Ave., Urbana, Illinois 61801. Currently with Qualcomm, Inc., 10555 Sorrento Valley Road, San Diego, California 92121.

[‡] Please send all correspondence to this author at: Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544 (poor@ivy.princeton.edu).

where $\mathbf{Q}_n$ is the covariance matrix of $\left[\delta^n X_k, \delta^{n-1} X_k, \dots, X_k\right]^T$ and $\tilde{\pi}_n = E(\nu_k^2)$.

We contrast this with the conventional autoregressive model

$$X_{k+n} + \sum_{j=1}^{n} a_{n,j} X_{k+n-j} = \epsilon_k,$$

in which the parameter vector $\underline{a}_n = [1, a_{n,1}, \dots, a_{n,n}]^T$ that minimizes $E(\epsilon_k^2)$ satisfies

$$\mathbf{R}_n \underline{a}_n = [\pi_n, 0, \dots, 0]^T, \tag{1.3}$$

where $\mathbf{R}_n$ is the Toeplitz covariance matrix of $[X_{k+n}, X_{k+n-1}, \dots, X_k]^T$ and $\pi_n = E(\epsilon_k^2)$. It is well known that (1.3) can be solved using the Levinson–Durbin algorithm [2], which requires $O(n^2)$ computations.

The matrix $\mathbf{Q}_n$ is not Toeplitz, but is related to the Toeplitz matrix $\mathbf{R}_n$ by

$$\mathbf{Q}_n = \mathbf{T}_n^T \mathbf{R}_n \mathbf{T}_n, \tag{1.4}$$

where $\mathbf{T}_n$ is an $(n+1) \times (n+1)$ lower triangular matrix whose l, jth element is given by

$$(\mathbf{T}_n)_{l,j} = \frac{(-1)^{l-j}}{\Delta^{n-j}} \binom{n-j}{l-j}, \qquad 0 \leq l,\ j \leq n. \tag{1.5}$$

(We follow the convention that the binomial coefficient $\binom{n}{j} = 0$ for $j < 0$ and $j > n$.) As a result, the vectors $\underline{a}_n$ and $\underline{\beta}_n$ that solve (1.3) and (1.2), respectively, are related by

$$\underline{\beta}_n = \Delta^{-n} \mathbf{T}_n^{-1} \underline{a}_n. \tag{1.6}$$

We define the predictor polynomials associated with $\mathbf{Q}_n$ and with the related Toeplitz matrix $\mathbf{R}_n$ by

$$A_n(z) = z^n + a_{n,1} z^{n-1} + \cdots + a_{n,n}$$

and

$$B_n(z) = z^n + \beta_{n,1} z^{n-1} + \cdots + \beta_{n,n},$$

respectively. Equation (1.6) is then equivalent to the polynomial relation

$$A_n(z) = \Delta^n B_n\left(\frac{z-1}{\Delta}\right).$$

Even though δ-based models are theoretically equivalent to shift-operator based models, in the fast sampling regime the normal equations associated with the δ models exhibit better numerical conditioning than do their shift-operator based counterparts. It has been shown in [9] that, even though the δ-based normal equations are not Toeplitz, their special structure can be exploited to derive an algorithm to solve them using only $O(n^2)$ computations, analogous to the Levinson–Durbin algorithm for the Toeplitz equations associated with the conventional model [6].

The Levinson-type algorithm described in [9] solves for the model parameters by computing the triangular factorization of the inverse of a covariance matrix of the differenced data. One handicap of that algorithm is that, since it requires the computation of n-dimensional inner products, it will require $O(n \log n)$ time even when n processors are used in parallel. In this paper we present an algorithm that computes the triangular factorization of the covariance matrix, without requiring any inner-product computations. This algorithm also requires only $O(n^2)$ computations overall, and has the added advantage that the time required is only $O(n)$ when n processors are used in parallel.

2. The algorithm. We derive an algorithm for finding the triangular factors of the $(n+1) \times (n+1)$ positive-definite matrix $\mathbf{Q}$ $(= \mathbf{Q}_n)$ defined in the Introduction.

The factorization is given by

$$\mathbf{Q} = \mathbf{LDL}^T, \tag{2.1}$$

where $\mathbf{L}$ is a lower triangular matrix with 1's on the diagonal, and $\mathbf{D}$ is a diagonal matrix. We actually compute a factorization of the form

$$\tilde{\mathbf{L}}\mathbf{Q} = \mathbf{U}, \tag{2.2}$$

where $\tilde{\mathbf{L}}$ is lower triangular and $\mathbf{U}$ is unit upper triangular. By comparing with (2.1), and by the uniqueness of the *LDU* decomposition [8], we see that $\mathbf{U} = \mathbf{L}^T$, and $\tilde{\mathbf{L}} = (\mathbf{LD})^{-1}$.

In order to compute the factorization efficiently, we use the following easily verifiable property of the matrix $\mathbf{Q}$:

$$\mathbf{Q}_{i,j} = -\frac{1}{\Delta}\left[\mathbf{Q}_{i,j+1} + \mathbf{Q}_{i+1,j}\right], \qquad 0 \leq i, j < n. \tag{2.3}$$

Equation (2.3) implies that the matrix $\mathbf{Q}$ has a *Toeplitz-like* structure as defined by Heinig and Rost [4], or a *displacement structure* as defined by Lev-Ari and Kailath [5]. The algorithm we derive is similar to one derived for Hankel matrices by Rissanen [7], and is also related to the fast Schur-type algorithm for Toeplitz matrices derived by Delsome and Ipsen in [1].

The factorization (2.2) can be expressed in terms of the vectors $\mathbf{z}^m = [z_0^m, \ldots, z_m^m]$, $m = 0, \ldots, n$, that solve the equations

$$\mathbf{z}^m \mathbf{Q}_{(m)} = [0, \ldots, 0, 1], \tag{2.4}$$

where $\mathbf{Q}_{(m)}$ is the $(m+1) \times (m+1)$ leading principal submatrix of $\mathbf{Q}$, i.e.,

$$\mathbf{Q}_{(m)} = \left[\mathbf{Q}_{i,j}\right]_{i,j=0}^{m}.$$

(Note that $\{-z_j^m / z_m^m\}_{j=1}^m$ are the coefficients of the regression of $\delta^{n-m} X_k$ on $\delta^n X_k$, $\delta^{n-1} X_k$, ..., $\delta^{n-m+1} X_k$.)

Since the $\mathbf{Q}_{(m)}$'s are nested, we can define a unit upper triangular matrix $\mathbf{S}$ with elements $\{s_j^i\}$ such that

$$\begin{bmatrix} z_0^0 & & & \\ z_0^1 & z_1^1 & & \\ \vdots & \vdots & \ddots & \\ z_0^n & \cdots & \cdots & z_n^n \end{bmatrix} \mathbf{Q} = \begin{bmatrix} 1 & s_1^0 & s_2^0 & \cdots & s_n^0 \\ & 1 & s_2^1 & \cdots & s_n^1 \\ & & \ddots & \cdots & \cdots \\ & & & \ddots & \cdots \\ & & & & 1 \end{bmatrix}, \tag{2.5}$$

where the blank spaces in the matrices represent zero elements. By comparing (2.5) with (2.2), we see that

$$\tilde{\mathbf{L}} = \begin{bmatrix} z_0^0 & & & \\ z_0^1 & z_1^1 & & \\ \vdots & \vdots & \ddots & \\ z_0^n & \cdots & \cdots & z_n^n \end{bmatrix}$$

and

$$\mathbf{U} = \begin{bmatrix} 1 & s_1^0 & s_2^0 & \cdots & s_n^0 \\ & 1 & s_2^1 & \cdots & s_n^1 \\ & & \ddots & \cdots & \cdots \\ & & & \ddots & \cdots \\ & & & & 1 \end{bmatrix}.$$

We now define an $(n+1) \times (2n+1)$ matrix $\tilde{\mathbf{Q}}$, formed by adding n columns to the right of $\mathbf{Q}$ in such a way as to preserve the structure given by (2.3), i.e.,

$$\tilde{\mathbf{Q}}_{i,j} = \mathbf{Q}_{i,j}, \qquad 0 \leq i, j \leq n,$$

$$\tilde{\mathbf{Q}}_{n,j} = -\Delta\tilde{\mathbf{Q}}_{n,j-1}, \qquad j = n+1, \ldots, 2n,$$

and

$$\tilde{\mathbf{Q}}_{i,j} = -\tilde{\mathbf{Q}}_{i+1,j-1} - \Delta\tilde{\mathbf{Q}}_{i,j-1}, \quad i = 0, 1, \ldots, n-1, \quad j = n+1, \ldots, 2n.$$

Equation (2.5) can then be augmented as follows:

$$\begin{bmatrix} z_0^0 & & & \\ z_0^1 & z_1^1 & & \\ \vdots & \vdots & \ddots & \\ z_0^n & \cdots & \cdots & z_n^n \end{bmatrix} \tilde{\mathbf{Q}} = \begin{bmatrix} 1 & s_1^0 & s_2^0 & \cdots & s_n^0 & \cdots & s_{2n}^0 \\ & 1 & s_2^1 & \cdots & s_n^1 & \cdots & s_{2n}^1 \\ & & \ddots & \cdots & \cdots & \cdots & \cdots \\ & & & \ddots & \cdots & \cdots & \cdots \\ & & & & 1 & \cdots & s_{2n}^n \end{bmatrix}. \tag{2.6}$$

We shall now obtain relations for recursively computing the rows of the matrices on the left- and right-hand sides of (2.6). In order to do this, we introduce the vectors $\mathbf{y}^m = [y_0^m, \ldots, y_m^m]$, $m = 0, \ldots, n$, that solve the equations

$$\mathbf{y}^m \mathbf{Q}_{(m)} = [1, 0, \ldots, 0]. \tag{2.7}$$

(Note that $\{-y_j^m / y_0^m\}_{j=1}^m$ are the coefficients of the regression of $\delta^n X_k$ on $\delta^{n-1} X_k$, $\ldots$, $\delta^{n-m} X_k$. For the case of $m = n$, by comparing with (1.2), we have $\mathbf{y}^n = \underline{\beta}_n y_0^n = \underline{\beta}_n / \tilde{\pi}_n$.)

We can define a matrix $\mathbf{R} = \{r_j^i\}$ with the structure shown in (2.8) such that the vectors $\mathbf{y}^m$ satisfy

$$\begin{bmatrix} y_0^0 & & & \\ y_0^1 & y_1^1 & & \\ \vdots & \vdots & \ddots & \\ y_0^n & \cdots & \cdots & y_n^n \end{bmatrix} \tilde{\mathbf{Q}} = \begin{bmatrix} 1 & r_1^0 & r_2^0 & \cdots & \cdots & \cdots & r_{2n}^0 \\ 1 & & r_2^1 & \cdots & \cdots & \cdots & r_{2n}^1 \\ \vdots & & & \ddots & \cdots & \cdots & \cdots \\ 1 & & & & r_{n+1}^n & \cdots & r_{2n}^n \end{bmatrix}. \tag{2.8}$$

For $m < n$, we define the row vector $\hat{\mathbf{y}}^m$ by

$$\hat{y}_i^m = \Delta y_i^m + y_{i-1}^m,$$

where we assume that $y_i^m = 0$ if $i < 0$ or $i > m$. Then, the jth element of

$$[\hat{\mathbf{y}}^m, 0, \ldots, 0]\tilde{\mathbf{Q}}$$

is given by

$$\begin{aligned}\sum_{i=0}^{m+1} \hat{y}_i^m \tilde{\mathbf{Q}}_{i,j} &= \Delta \sum_{i=0}^{m} y_i^m \tilde{\mathbf{Q}}_{i,j} + \sum_{i=1}^{m+1} y_{i-1}^m \tilde{\mathbf{Q}}_{i,j} \\ &= \Delta \sum_{i=0}^{m} y_i^m \tilde{\mathbf{Q}}_{i,j} + \sum_{i=0}^{m} y_i^m \tilde{\mathbf{Q}}_{i+1,j} \\ &= \Delta \sum_{i=0}^{m} y_i^m \tilde{\mathbf{Q}}_{i,j} + \sum_{i=0}^{m} y_i^m \left[-\tilde{\mathbf{Q}}_{i,j+1} - \Delta \tilde{\mathbf{Q}}_{i,j}\right] \\ &= -\sum_{i=0}^{m} y_i^m \tilde{\mathbf{Q}}_{i,j+1} \quad \text{if } j < 2n.\end{aligned}$$

Therefore,

$$\left[\hat{y}_0^m, \ldots, \hat{y}_{m+1}^m, 0, \ldots, 0\right] \tilde{\mathbf{Q}} = \left[0, \ldots, 0, -r_{m+1}^m, \ldots, -r_{2n}^m, x\right], \tag{2.9}$$

where the right-hand side has m leading zeros. By equating the mth row on either side of (2.6), we get

$$\left[z_0^m, \ldots, z_m^m, 0, \ldots, 0\right] \tilde{\mathbf{Q}} = \left[0, \ldots, 0, 1, s_{m+1}^m, \ldots, s_{2n}^m\right].$$

Hence, using the condition that the first $m+1$ elements of

$$\left[z_0^{m+1}, \ldots, z_{m+1}^{m+1}, 0, \ldots, 0\right] \tilde{\mathbf{Q}}$$

are zero, we get

$$z_j^{m+1} = (y_{j-1}^m + \Delta y_j^m + r_{m+1}^m z_j^m)/K_m, \qquad j = 0, \ldots, m+1, \tag{2.10}$$

for some constant K_m. From this relationship, we get the following recursion for the right-hand side of (2.6):

$$s_j^{m+1} = (r_{m+1}^m s_j^m - r_{j+1}^m)/K_m. \tag{2.11}$$

The condition that $s_m^m = 1$ yields the following expression for K_m:

$$K_m = r_{m+1}^m s_{m+1}^m - r_{m+2}^m. \tag{2.12}$$

Now we have

$$\left[y_0^m, \ldots, y_m^m, 0, \ldots, 0\right] \tilde{\mathbf{Q}} = \left[1, 0, \ldots, 0, r_{m+1}^m, \ldots, r_{2n}^m\right]$$

and

$$\left[z_0^{m+1}, \ldots, z_{m+1}^{m+1}, 0, \ldots, 0\right] \tilde{\mathbf{Q}} = \left[0, \ldots, 0, 1, s_{m+2}^{m+1}, \ldots, s_{2n}^m\right].$$

From these, we have the recursions

$$y_j^{m+1} = (y_j^m - r_{m+1}^m z_j^{m+1}) \tag{2.13}$$

and

$$r_j^{m+1} = r_j^m - r_{m+1}^m s_j^{m+1}. \tag{2.14}$$

The algorithm is initialized by setting

$$z_0^0 = y_0^0 = \frac{1}{\mathbf{Q}_{0,0}} \tag{2.15}$$

and

$$s_j^0 = r_j^0 = \mathbf{Q}_{0,j}/\mathbf{Q}_{0,0}, j = 0, \ldots, 2n. \tag{2.16}$$

In order for the recursions (2.10) and (2.11) to be valid, we need to show that $K_m \neq 0$. By setting $j = m + 1$ in (2.10), we get

$$K_m = y_m^m / z_{m+1}^{m+1}.$$

Since $\mathbf{Q}$, and hence $\mathbf{Q}_{(m+1)}$, are positive definite, z_{m+1}^{m+1} and y_0^m are positive. If $B_{(m)}(z)$ and $A_{(m)}(z)$ are the predictor polynomials associated with $\mathbf{Q}_{(m)}$ and with its related Toeplitz matrix, respectively, then

$$y_m^m = y_0^m B_{(m)}(0) = y_0^m \Delta^{-n} A_{(m)}(1) > 0,$$

since $A_{(m)}(z)$ is a stable predictor polynomial. Thus, we have shown that K_m is positive.

The LDL^T factorization of $\mathbf{Q}$ can be computed using (2.11), (2.14), and (2.12). These equations can also be used in conjunction with (2.10) and (2.13) to compute the UDU^T factorization of $\mathbf{Q}^{-1}$.

It should be noted that, because of the number x being shifted in from the right in (2.9), (2.11) holds only for $j \leq 2n - m$. However, this does not affect our calculations since we are interested only in computing s_j^i for $j \leq n$. This explains the need for forming the extended matrix $\tilde{\mathbf{Q}}$; the extension ensures that the "error propagation" from the right does not affect the variables in which we are interested.

3. Conclusions. In this paper, we have presented an $O(n^2)$ algorithm for computing the triangular factors of the covariance matrix of a random vector whose elements are successive differences of a stationary time series. The derivation has been carried out by exploiting the Toeplitz-like nature of the matrix. The algorithm does not require the computation of any n-dimensional inner products, and, as a result, can be parallelized to yield $O(n)$ time complexity by using n processors in parallel. This algorithm has applications in modeling time series using difference operators. Such models have been motivated by the need to avoid the problem of numerical ill-conditioning of the Toeplitz normal equations associated with the conventional autoregressive model when applied to rapidly sampled processes. A number of such applications are discussed in [3].

REFERENCES

[1] J.-M. Delsome and I. Ipsen, *Parallel solution of symmetric positive definite systems with hyperbolic rotations*, Linear Algebra Appl., 77 (1986), pp. 75–111.

[2] G. H. Golub and C. F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[3] G. C. Goodwin, R. H. Middleton, and H. V. Poor, *High-speed digital signal processing and control*, Proc. of the IEEE, 80, 1992, pp. 240–259.

[4] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like Matrices and Operators*, Birkhäuser, Basel, 1984.

[5] H. Lev-Ari and T. Kailath, *Triangular factorization of structured hermitian matrices*, in I. Schur Methods in Operator Theory and Signal Processing, I. Gohberg, ed., Birkhäuser, Boston, 1986, pp. 301–324.

[6] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, New York, 1988.

[7] J. Rissanen, *Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with application to factoring positive matrix polynomials*, Math. Comput., 27 (1973), pp. 147–154.

[8] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[9] R. Vijayan, H. V. Poor, J. B. Moore, and G. C. Goodwin, *A Levinson-type algorithm for modeling fast-sampled data*, IEEE Trans. Automatic Control, AC-36, 1991, pp. 314–321.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1103–1111, October 1993

014

ON THE SMITH NORMAL FORM OF STRUCTURED POLYNOMIAL MATRICES, II*

KAZUO MUROTA†

Abstract. The Smith normal form of a polynomial matrix $D(s) = Q(s) + T(s)$ is investigated, where $D(s)$ is structured in the sense that (i) the coefficients of the entries of $Q(s)$ belong to a field $\boldsymbol{K}$ and (ii) the nonzero coefficients of the entries of $T(s)$ are algebraically independent parameters over $\boldsymbol{K}$. It is shown that all the invariant polynomials except for the last do not contain the system parameters (= coefficients of $T(s)$) and that the last invariant polynomial is expressed in terms of the combinatorial canonical form (CCF) of a layered mixed matrix associated with $D(s)$. This implies that the generic dependency of the invariant polynomials on the system parameters can be computed by means of a matroid-theoretic algorithm that involves arithmetic operations in $\boldsymbol{K}(s)$.

Key words. Smith normal form, structured polynomial matrix, mixed matrix, combinatorial canonical form (CCF)

AMS subject classifications. 15A21, 15A54, 05C50, 93B

1. Introduction. A previous paper [6] by the author investigated the Smith normal form (e.g., [1], [9]) of a polynomial matrix $D(s) = Q_D(s) + T_D(s)$ with coefficients from a field $\boldsymbol{F}$ that are structured in the following sense:

(A1) The coefficients of the entries of $Q_D(s)$ belong to a subfield $\boldsymbol{K}$ of $\boldsymbol{F}$.

(A2) The nonzero coefficients $\mathcal{T}$ ($\subseteq \boldsymbol{F}$) of the entries of $T_D(s)$ are algebraically independent over $\boldsymbol{K}$.

(A3) Every minor of $Q_D(s)$ is a monomial in s.

See [3], [6] for the motivation for such polynomial matrices. It has been shown in [6] that the invariant polynomials of such a matrix, except for the last one, are monomials in s and that the last invariant polynomial can be expressed in terms of the combinatorial canonical form (CCF) of a layered mixed matrix (LM-matrix) associated with $D(s)$. On the basis of these results an efficient matroid-theoretic algorithm has been proposed for computing the Smith normal form.

In this paper we consider the Smith normal form of a polynomial matrix $D(s)$ having the two properties (A1) and (A2) only. We are mainly interested in how the invariant polynomials depend generically on the system parameters (= coefficients of $T(s)$).

The main theorems, which are given in §2, state that if $D(s)$ satisfies (A1) and (A2), all the invariant polynomials of $D(s)$ except for the last are polynomials in s free from the system parameters and the last invariant polynomial can be expressed in terms of the combinatorial canonical form of an LM-matrix associated with $D(s)$. This implies that the generic dependency of the Smith normal form on the system parameters can be computed by means of a matroid-theoretic algorithm that involves arithmetic operations in the field of rational functions $\boldsymbol{K}(s)$.

The following examples illustrate the class of matrices and the problem considered in this paper.

* Received by the editors July 15, 1991; accepted for publication (in revised form) February 3, 1993. This paper was presented at the International Symposium on the Mathematical Theory of Networks and Systems held in Kobe, Japan, June 17–21, 1991.

† Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606, Japan (murota@kurims.kyoto-u.ac.jp).

Example 1.1. Consider a 2×2 matrix

$$D(s) = \begin{pmatrix} s+p_1 & s+p_3 \\ 0 & p_2 s+1 \end{pmatrix}.$$

Here $\mathcal{T} = \{p_1, p_2, p_3\}$ is the set of (nonzero) system parameters. This matrix is expressed as $D(s) = Q_D(s) + T_D(s)$ (where $\boldsymbol{K} = \mathbf{Q}$, $\boldsymbol{F} = \mathbf{Q}(p_1, p_2, p_3)$) with

$$Q_D(s) = \begin{pmatrix} s & s \\ 0 & 1 \end{pmatrix}, \quad T_D(s) = \begin{pmatrix} p_1 & p_3 \\ 0 & p_2 s \end{pmatrix}.$$

The Smith normal form of $D(s)$ as a matrix over $\boldsymbol{F}[s]$ is given by

$$\Sigma(s) = \begin{cases} \mathrm{diag}\,[s+p_1,\ s+p_1] & \text{if } p_1 = 1/p_2 = p_3, \\ \mathrm{diag}\,[1,\ (s+p_1)(s+1/p_2)] & \text{otherwise.} \end{cases}$$

This shows that generically speaking, or if there is no algebraic relation among the system parameters $\mathcal{T}$, the first invariant polynomial is free from $\mathcal{T}$ and the second depends on (p_1, p_2) and not on p_3. In this paper we are concerned with such generic dependency of the Smith normal form on the system parameters.

Example 1.2. Consider a 5×5 matrix

$$(1) \qquad D(s) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} 0 & 0 & 1+p_1 s & 3s^2+3s & p_2 \\ s^2+s & s+1 & 1 & 0 & p_3+p_4 s \\ 2s^3+2s^2 & 2s^2+2s & 2s & 0 & p_5 s \\ 0 & 0 & 0 & s^2-1 & p_6 \\ 2s^4+2s^3 & 2s^3+2s^2 & 2s^2 & 0 & s+p_7 s^2 \end{pmatrix} \end{array}.$$

The row set and the column set of D are respectively denoted as $\mathrm{Row}(D) = \{w_1, w_2, w_3, w_4, w_5\}$, $\mathrm{Col}(D) = \{x_1, x_2, x_3, x_4, x_5\}$. Here $\mathcal{T} = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$ is the set of algebraically independent system parameters. This matrix is expressed as $D(s) = Q_D(s) + T_D(s)$ (for $\boldsymbol{K} = \mathbf{Q}$) with

$$(2) \qquad Q_D(s) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 3s^2+3s & 0 \\ s^2+s & s+1 & 1 & 0 & 0 \\ 2s^3+2s^2 & 2s^2+2s & 2s & 0 & 0 \\ 0 & 0 & 0 & s^2-1 & 0 \\ 2s^4+2s^3 & 2s^3+2s^2 & 2s^2 & 0 & s \end{pmatrix} \end{array},$$

$$(3) \qquad T_D(s) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} 0 & 0 & p_1 s & 0 & p_2 \\ 0 & 0 & 0 & 0 & p_3+p_4 s \\ 0 & 0 & 0 & 0 & p_5 s \\ 0 & 0 & 0 & 0 & p_6 \\ 0 & 0 & 0 & 0 & p_7 s^2 \end{pmatrix} \end{array}.$$

The Smith normal form of $D(s)$ as a matrix over $\boldsymbol{F}[s]$ is

$$(4) \qquad \Sigma(s) = \mathrm{diag}\,[1,\ 1,\ s+1,\ s(s-1)(s+1)(s+1/p_1),\ 0],$$

so long as $\mathcal{T}$ is algebraically independent. Note that the first three invariant polynomials, namely, 1, 1, $s+1$, do not contain the parameters p_i and that the last one depends only on p_1. The main objective of this paper is to identify the parameters that appear generically in the Smith normal form.

2. Results. The results of this paper are presented as Theorems 2.1–2.3 and are illustrated in Example 2.1. Their proofs are postponed until the next section. We use the terminology introduced in the previous paper [6].

Let $D(s)$ be an $m \times n$ polynomial matrix in indeterminate s over a field $\boldsymbol{F}$ represented as

$$D(s) = Q_D(s) + T_D(s), \tag{5}$$

with properties (A1) and (A2) in §1. This implies that $D(s)$ is a mixed matrix (see [6]–[8]) with respect to the rational function field $\boldsymbol{K}(s)$.

Let $d_k(s)$ denote the kth determinantal divisor of $D(s)$ over $\boldsymbol{F}$ for $k = 1, \ldots, r$, where $r = \operatorname{rank} D(s)$. Then the kth invariant polynomial $e_k(s)$ of $D(s)$ is expressed as

$$e_k(s) = d_k(s)/d_{k-1}(s) \quad \text{for } k = 1, \ldots, r, \tag{6}$$

where $d_0(s) = 1$ by convention. The Smith normal form of $D(s)$ as a matrix over $\boldsymbol{F}[s]$ is given by

$$\Sigma(s) = \operatorname{diag}[e_1(s), \ldots, e_r(s), 0, \ldots, 0].$$

We choose $d_k(s)$ and $e_k(s)$ to be monic in $\boldsymbol{F}[s]$. Note that the coefficients of $d_k(s)$ and $e_k(s)$ are, in general, rational functions in $\mathcal{T}$ over $\boldsymbol{K}$.

The main objective of this paper is to show that the Smith form $\Sigma(s)$ of $D(s)$ has special properties, as stated in Theorems 2.1 and 2.2. The former refers to $e_k(s)$ for $k = 1, \ldots, r-1$, whereas the latter refers to $e_r(s)$. The following arguments are very similar to those of the previous paper [6], in which condition (A3) of §1 is further imposed.

THEOREM 2.1. *For $k = 1, \ldots, r-1$ $(= \operatorname{rank} D(s) - 1)$ the kth monic determinantal divisors and the kth monic invariant polynomials of $D(s)$ contain no elements of $\mathcal{T}$. That is, $d_k(s) \in \boldsymbol{K}[s]$, $e_k(s) \in \boldsymbol{K}[s]$ for $k = 1, \ldots, r-1$.*

To determine the last invariant polynomial $e_r(s)$ we associate with $D(s)$ an augmented $(2m) \times (m+n)$ matrix

$$\tilde{D}(s) = \tilde{D}(s;t) = \begin{pmatrix} I_m & Q_D(s) \\ -\operatorname{diag}[t_1, \ldots, t_m] & T_D(s) \end{pmatrix} = \begin{pmatrix} \tilde{Q}(s) \\ \tilde{T}(s;t) \end{pmatrix} \tag{7}$$

with new indeterminates $t = (t_1, \ldots, t_m)$ in $\boldsymbol{F}$, where

$$\tilde{Q}(s) = (I_m \mid Q_D(s)), \quad \tilde{T}(s;t) = (-\operatorname{diag}[t_1, \ldots, t_m] \mid T_D(s)).$$

It is not difficult to see that $\tilde{D}(s;1) = \tilde{D}(s;t)|_{t_1 = \cdots = t_m = 1}$ is equivalent, as a polynomial matrix in s over $\boldsymbol{F}$, to $\operatorname{diag}[I_m, D(s)]$, and hence the Smith form $\Sigma(s)$ of $D(s)$ is embedded in the Smith form of $\tilde{D}(s;1)$ as

$$\begin{pmatrix} I_m & O \\ O & \Sigma(s) \end{pmatrix}. \tag{8}$$

In particular, $e_r(s)$ is equal to the last invariant polynomial of $\tilde{D}(s;1)$.

Since $\tilde{D}(s;t)$ is an LM-matrix with respect to $\boldsymbol{K}(s)$, we can talk of its CCF, say, $\bar{D}(s;t)$. See [6, §2] for CCF. Let $\{\bar{D}_l(s;t) \mid l = 0, 1, \ldots, b, \infty\}$ denote the family of its irreducible diagonal blocks, where $\bar{D}_0(s;t)$ and $\bar{D}_\infty(s;t)$ are the horizontal and the vertical tails, respectively, and $\bar{D}_l(s;t)$ $(l = 1, \ldots, b)$ are square irreducible blocks. As mentioned in [6, Remark 2.1], there is some indeterminacy in the numerical values of the entries of $\bar{D}_l(s;t)$, although the partitions of the rows and the columns are uniquely determined. The following theorem states that $e_r(s)$ is characterized by those diagonal blocks when they are appropriately chosen; in particular, the statement of the theorem presupposes that we can choose $\bar{D}_l(s;t)$ in such a way that the diagonal blocks are polynomial matrices in s and $\mathcal{T}$ over $\boldsymbol{K}$. See Lemma 3.2 in §3.

THEOREM 2.2. *The rth monic determinantal divisor and the rth monic invariant polynomial of $D(s)$ as a matrix over $\boldsymbol{F}[s]$ can be expressed in terms of the irreducible blocks $\{\bar{D}_l(s;t) \mid l = 0, 1, \ldots, b, \infty\}$ of an appropriately chosen* CCF $\bar{D}(s;t)$ *of* $\tilde{D}(s;t)$ *as follows, where* $r = \operatorname{rank} D(s)$.

$$d_r(s) = \alpha_r \cdot \bar{d}_r(s) \cdot \prod_{l=1}^{b} \det \bar{D}_l(s;1), \tag{9}$$

$$e_r(s) = \alpha_r \cdot \bar{e}_r(s) \cdot \prod_{l=1}^{b} \det \bar{D}_l(s;1) \tag{10}$$

for some $\alpha_r \in \boldsymbol{F} - \{0\}$, $\bar{d}_r(s) \in \boldsymbol{K}[s]$, and $\bar{e}_r(s) \in \boldsymbol{K}[s]$. Moreover, the monic determinantal divisor $\bar{d}^{0}_r(s)$ (respectively, $\bar{d}^{\infty}_r(s)$) $\in \boldsymbol{F}[s]$ of the largest order of the horizontal (respectively, vertical) tail $\bar{D}_0(s;t)$ (respectively, $\bar{D}_\infty(s;t)$) belongs to $\boldsymbol{K}[s]$, and $\bar{d}_r(s) = \bar{d}^{0}_r(s) \cdot \bar{d}^{\infty}_r(s)$.

Remark 2.1. The previous paper [6] treats the special case in which $D(s)$ possesses the additional property (A3): Every minor of $Q_D(s)$ is a monomial in s. For this case we can further restrict $d_k(s)$, $e_k(s)$ $(k = 1, \ldots, r-1)$ in Theorem 2.1 and $\bar{d}_r(s)$, $\bar{e}_r(s)$ in Theorem 2.2 to monomials.

Theorem 2.2 shows that parameters that are not contained in the square diagonal blocks $\{\bar{D}_l(s;t) \mid l = 1, \ldots, b\}$ of the CCF have no generic influence on the Smith form. The converse is also true, as stated below, since each parameter contained in a square diagonal block $\bar{D}_l(s;t)$ does appear in its determinant by the irreducibility of the block (see [4, Thm. 3.5], [5, Lemma 5.1], and [7]).

THEOREM 2.3. *The rth monic invariant polynomial of $D(s)$ generically depends on exactly those parameters in $\mathcal{T}$ that are contained in the square irreducible blocks $\{\bar{D}_l(s;t) \mid l = 1, \ldots, b\}$ of the* CCF *of* $\tilde{D}(s;t)$.

Example 2.1. Theorems 2.1–2.3 are illustrated here for matrix (1) of Example 1.2. The augmented LM-matrix $\tilde{D}(s;t)$ of (7) is given by

$$\tilde{Q}(s) = \begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array}\left(\begin{array}{ccccc|cccccc} w_1 & w_2 & w_3 & w_4 & w_5 & x_1 & x_2 & x_3 & x_4 & x_5 \\ 1 & & & & & 0 & 0 & 1 & 3s^2+3s & 0 \\ & 1 & & & & s^2+s & s+1 & 1 & 0 & 0 \\ & & 1 & & & 2s^3+2s^2 & 2s^2+2s & 2s & 0 & 0 \\ & & & 1 & & 0 & 0 & 0 & s^2-1 & 0 \\ & & & & 1 & 2s^4+2s^3 & 2s^3+2s^2 & 2s^2 & 0 & s \end{array}\right),$$

$$\tilde{T}(s;t) = \left(\begin{array}{cccccccccc} w_1 & w_2 & w_3 & w_4 & w_5 & x_1 & x_2 & x_3 & x_4 & x_5 \\ -t_1 & & & & & 0 & 0 & p_1 s & 0 & p_2 \\ & -t_2 & & & & 0 & 0 & 0 & 0 & p_3 + p_4 s \\ & & -t_3 & & & 0 & 0 & 0 & 0 & p_5 s \\ & & & -t_4 & & 0 & 0 & 0 & 0 & p_6 \\ & & & & -t_5 & 0 & 0 & 0 & 0 & p_7 s^2 \end{array}\right).$$

By means of the admissible transformation (see [6, §2]),

$$\bar{D}(s;t) = P_{\mathrm{r}} \begin{pmatrix} S(s) & O \\ O & I \end{pmatrix} \begin{pmatrix} \tilde{Q}(s) \\ \tilde{T}(s;t) \end{pmatrix} P_{\mathrm{c}},$$

where P_{r} and P_{c} are permutations and

$$S(s) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \!\!\begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & w_4 & w_5 \end{array} \\ \left(\begin{array}{ccccc} 1 & & & \frac{-3s}{s-1} & \\ & 1 & & & \\ & -2s & 1 & & \\ & & & 1 & \\ & -2s^2 & & & 1 \end{array}\right) \end{array}, \tag{11}$$

we obtain the following CCF of $\tilde{D}(s;t)$:

$$\bar{D}(s;t) = \left(\begin{array}{cccccccccc} x_1 & x_2 & w_1 & x_3 & x_4 & w_4 & w_3 & w_5 & w_2 & x_5 \\ s^2+s & s+1 & & 1 & & & & & 1 & \\ & & 1 & 1 & & \frac{-3s}{s-1} & & & & \\ & & -t_1 & p_1s & & & & & & p_2 \\ & & & & s^2-1 & 1 & & & & \\ & & & & & -t_4 & & & & p_6 \\ & & & & & & 1 & 0 & -2s & 0 \\ & & & & & & 0 & 1 & -2s^2 & s \\ & & & & & & -t_3 & 0 & 0 & p_5s \\ & & & & & & 0 & -t_5 & 0 & p_7s^2 \\ & & & & & & 0 & 0 & -t_2 & p_3+p_4s \end{array}\right).$$

We have nonempty tails,

$$\bar{D}_0(s;t) = \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{pmatrix} s^2+s & s+1 \end{pmatrix} \end{array},$$

$$\bar{D}_\infty(s;t) = \begin{array}{c} \begin{array}{cccc} w_3 & w_5 & w_2 & x_5 \end{array} \\ \begin{pmatrix} 1 & 0 & -2s & 0 \\ 0 & 1 & -2s^2 & s \\ -t_3 & 0 & 0 & p_5s \\ 0 & -t_5 & 0 & p_7s^2 \\ 0 & 0 & -t_2 & p_3+p_4s \end{pmatrix} \end{array},$$

and $b = 3$ square irreducible diagonal blocks:

$$\bar{D}_1(s;t) = \begin{array}{c} \begin{array}{cc} w_1 & x_3 \end{array} \\ \begin{pmatrix} 1 & 1 \\ -t_1 & p_1s \end{pmatrix} \end{array}, \quad \bar{D}_2(s;t) = \begin{array}{c} x_4 \\ \begin{pmatrix} s^2-1 \end{pmatrix} \end{array}, \quad \bar{D}_3(s;t) = \begin{array}{c} w_4 \\ \begin{pmatrix} -t_4 \end{pmatrix} \end{array}.$$

Note that the diagonal blocks are polynomial matrices in s and $\mathcal{T}$ over $\boldsymbol{K} = \mathbf{Q}$, whereas a fraction $-3s/(s-1)$ is contained in an off-diagonal block.

The CCF reveals that $r = \operatorname{rank} D(s) = 4$ (< 5). Then according to Theorem 2.2 we see that

$$d_4(s) = \alpha_4 \cdot \bar{d}_4(s) \cdot (p_1s+1) \cdot (s^2-1) \cdot (-1)$$

for some monic $\bar{d}_4(s) \in \boldsymbol{K}[s]$ and a normalizing factor $\alpha_4 = -1/p_1 \in \boldsymbol{F}$. The determinantal divisors of the tails $\bar{D}_0(s;t)$ and $\bar{D}_\infty(s;t)$ are $\bar{d}_4^0(s) = s+1$ and $\bar{d}_4^\infty(s) = s$, respectively, from which it follows that $\bar{d}_4(s) = s(s+1)$. The other determinantal divisors are found as $d_1(s) = d_2(s) = 1$, $d_3(s) = s+1$. Thus we obtain the Smith form $\Sigma(s)$ of (4). Note that p_1 is the only member of $\mathcal{T}$ that is contained in the square irreducible blocks $\bar{D}_l(s;t)$.

3. Proofs.

The proofs for Theorems 2.1 and 2.2 are almost parallel to the arguments in [6].

A minor (subdeterminant) of $D(s) = D(s, \mathcal{T})$, which can be regarded as a matrix over $\boldsymbol{K}[s, \mathcal{T}]$, is a polynomial in s and $\mathcal{T}$ over $\boldsymbol{K}$. Let $d_k^*(s, \mathcal{T}) \in \boldsymbol{K}[s, \mathcal{T}]$ denote the kth determinantal divisor of D, i.e., the greatest common divisor of all minors of order k in D as polynomials in $(s, \mathcal{T})$ over $\boldsymbol{K}$. To prove Theorem 2.1 it suffices to show the following.

LEMMA 3.1. *$d_{r-1}^*(s, \mathcal{T}) \in \boldsymbol{K}[s]$, i.e., no $p \in \mathcal{T}$ appears in d_{r-1}^*.*

Proof. Since $r = \operatorname{rank} D$, there exists a nonsingular submatrix $D[I, J]$ with $|I| = |J| = r$. For $p \in \mathcal{T}$ let (i, j) denote the position at which p appears in D. If p does not appear in $\delta = \det D[I, J]$ $(\neq 0)$, then d_{r-1}^* is free from p since d_{r-1}^* divides δ. If p does appear in δ, then $i \in I$ and $j \in J$ and furthermore $\delta' = \det D[I - i, J - j] \neq 0$. Obviously, δ' does not contain p, and hence d_{r-1}^* is free from p since d_{r-1}^* divides δ'. □

We now turn to the proof of Theorem 2.2. The matrix $\tilde{D}(s;t)$ of (7) plays the primary role here, since the Smith form of $D(s)$ is obtained from that of $\tilde{D}(s;1)$, as noted in (8). Recall that $\tilde{D}(s;t)$ is an LM-matrix with respect to $\boldsymbol{K}(s)$.

We denote by $\{C_0; C_1, \ldots, C_b; C_\infty\}$ and $\{R_0; R_1, \ldots, R_b; R_\infty\}$ the partitions of the column set and the row set of the CCF, say, $\bar{D}(s) = \bar{D}(s;t)$, of $\tilde{D}(s;t)$ as an LM-matrix with respect to $\boldsymbol{K}(s)$. Then, for some nonsingular rational function matrix $S(s)$ over $\boldsymbol{K}(s)$ we have

$$\bar{D}(s) = \bar{D}(s;t) = P_{\mathrm{r}} \begin{pmatrix} S(s) & O \\ O & I \end{pmatrix} \begin{pmatrix} \tilde{Q}(s) \\ \tilde{T}(s;t) \end{pmatrix} P_{\mathrm{c}}. \tag{12}$$

In general, the transformation (12) does not qualify as an equivalence transformation of $\tilde{D}(s;t)$ as a polynomial matrix in s, since $S(s)$ can involve rational functions of s; see the matrix $S(s)$ of (11) in Example 2.1. The following lemma claims, however, that we may restrict ourselves to a unimodular transformation of the form (12) if we do not care about the upper off-diagonal blocks in the transformed matrix; see Example 3.1 below.

LEMMA 3.2. *There exists a unimodular polynomial matrix $U(s)$ over $\boldsymbol{K}[s]$ (i.e., $\det U(s) \in \boldsymbol{K} - \{0\}$) such that*

$$\hat{D}(s;t) = P_{\mathrm{r}} \begin{pmatrix} U(s) & O \\ O & I \end{pmatrix} \tilde{D}(s;t) P_{\mathrm{c}} \tag{13}$$

is in the same block-triangular form as a CCF *$\bar{D}(s;t)$ of $\tilde{D}(s;t)$ and such that the diagonal blocks of $\hat{D}(s;t)$ coincide with those of $\bar{D}(s;t)$; i.e.,*

$$\hat{D}(s;t)[R_k, C_l] = O \quad \textit{if } 0 \leq l < k \leq \infty, \tag{14}$$

$$\hat{D}(s;t)[R_k, C_k] = \bar{D}(s;t)[R_k, C_k] \quad \textit{for } k = 0, l, \ldots, b, \infty. \tag{15}$$

Proof. From the construction of the CCF (see [3], [6]–[8]) it suffices to note the following fundamental fact. □

LEMMA 3.3. *For a matrix $A(s)$ over $\boldsymbol{K}[s]$ there exists a unimodular matrix $U(s)$ over $\boldsymbol{K}[s]$ such that*

$$U(s)A(s) = \begin{pmatrix} A_0(s) \\ O \end{pmatrix}$$

with $\operatorname{rank} A(s) = \operatorname{rank} A_0(s) = |\mathrm{Row}(A_0)|$.

Proof. This follows, e.g., from the result on the Hermite normal form [9]. □

Lemma 3.2 shows that $\tilde{D}(s;t)$ and $\hat{D}(s;t)$ share the same Smith form since they are connected by a unimodular transformation. On the other hand, (8) shows that the Smith form $\Sigma(s)$ of $D(s)$ is embedded in the Smith form of $\tilde{D}(s;1)$. The following lemma claims that $\tilde{D}(s;t)$ and $\tilde{D}(s;1)$ have essentially identical Smith forms. We write $\tilde{D}(s;t,\mathcal{T})$ for $\tilde{D}(s;t)$ to explicitly indicate its dependence on the coefficients $\mathcal{T}$ in $T_D(s)$.

LEMMA 3.4 (see [6]). *The Smith form of $\tilde{D}(s;1,\mathcal{T})$ is obtained from that of $\tilde{D}(s;t,\mathcal{T})$ by setting $t_1 = \cdots = t_m = 1$, and, conversely, the Smith form of $\tilde{D}(s;t,\mathcal{T})$ is obtained from that of $\tilde{D}(s;1,\mathcal{T})$ by replacing $\tilde{t}$ with $\tilde{t}/t_i$ if $\tilde{t} \in \mathcal{T}$ is contained in row i.*

Thus we have established a link among the Smith forms, which may be schematically displayed as

$$D(s) \overset{(8)}{\longleftrightarrow} \tilde{D}(s;1) \overset{\text{Lemma 3.4}}{\longleftrightarrow} \tilde{D}(s;t) \overset{\text{Lemma 3.2}}{\longleftrightarrow} \hat{D}(s;t). \tag{16}$$

This allows us to concentrate on the Smith form of $\hat{D}(s;t)$. Regarding $\hat{D}(s;t) = \hat{D}(s;t,\mathcal{T})$ as a matrix over the ring $\boldsymbol{K}[s,t,\mathcal{T}]$, we denote by $\hat{d}_k(s;t)$ $(\in \boldsymbol{K}[s,t,\mathcal{T}])$ the kth determinantal divisor of $\hat{D}(s;t)$ for $k = 1,\ldots,r+m$. Then

$$d_k(s) = \alpha_k \cdot \hat{d}_{k+m}(s;1) \text{ for } k = 1,\ldots,r, \tag{17}$$

where α_k $(\in \boldsymbol{K}(\mathcal{T}) \subseteq \boldsymbol{F})$ is introduced since $d_k(s)$ was defined to be a monic polynomial in $\boldsymbol{F}[s]$. Since $\hat{D}$ is in the block-triangular form (14) with full-rank diagonal blocks (see (15) and [6, Thm. 2.3(c)]), we have $r + m = \operatorname{rank}\hat{D} = \sum_{l=0}^{b} |R_l| + |C_\infty|$, and therefore a nonvanishing minor of $\hat{D}$ of order $r+m$ is expressed as

$$\begin{aligned} &\det \hat{D}[R_0,J] \cdot \det \hat{D}[I,C_\infty] \cdot \prod_{l=1}^{b} \det \hat{D}[R_l,C_l] \\ &\quad = \det \bar{D}[R_0,J] \cdot \det \bar{D}[I,C_\infty] \cdot \prod_{l=1}^{b} \det \bar{D}[R_l,C_l] \end{aligned} \tag{18}$$

for some $J \subseteq C_0$ and $I \subseteq R_\infty$. Then Theorem 2.2 follows from (17) and (18) and the lemma below, where $\gcd_{\boldsymbol{K}[s,t,\mathcal{T}]}\{\cdot\}$ denotes the greatest common divisor in the ring $\boldsymbol{K}[s,t,\mathcal{T}]$.

LEMMA 3.5.

$$\gcd{}_{\boldsymbol{K}[s,t,\mathcal{T}]}\{\det \bar{D}[R_0,J] \mid |J| = |R_0|,\ J \subseteq C_0\} = \alpha_0(t,\mathcal{T}) \cdot \bar{d}^0(s),$$
$$\gcd{}_{\boldsymbol{K}[s,t,\mathcal{T}]}\{\det \bar{D}[I,C_\infty] \mid |I| = |C_\infty|,\ I \subseteq R_\infty\} = \alpha_\infty(t,\mathcal{T}) \cdot \bar{d}^\infty(s),$$

where $\alpha_0(t,\mathcal{T}), \alpha_\infty(t,\mathcal{T}) \in \boldsymbol{K}[t,\mathcal{T}]$, *and* $\bar{d}^0(s), \bar{d}^\infty(s) \in \boldsymbol{K}[s]$.

Proof. Regarding $\bar{D}[R_0, C_0]$ as an irreducible LM-matrix with respect to $\boldsymbol{K}(s)$, we obtain the first identity from [6, Thm. 2.5(b)]. The second follows similarly from [6, Thm. 2.5(c)]. □

Example 3.1. Lemma 3.2 is illustrated here for the matrix $D(s)$ of (1) in Example 1.2. Using a unimodular polynomial matrix

$$U(s) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \!\!\begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & w_4 & w_5 \end{array} \\ \left(\begin{array}{ccccc} 1 & & & & \\ & 1 & & & \\ & -2s & 1 & & \\ & & & 1 & \\ & -2s^2 & & & 1 \end{array}\right) \end{array}$$

in (13), we obtain the block-triangular matrix

$$\hat{D}(s;t) = \begin{array}{c} \begin{array}{cccccccccc} x_1 & x_2 & w_1 & x_3 & x_4 & w_4 & w_3 & w_5 & w_2 & x_5 \end{array} \\ \left(\begin{array}{cc|cc|c|c|cccc} s^2+s & s+1 & & 1 & & & & & 1 & \\ \hline & & 1 & 1 & 3s^2+3s & & & & & \\ & & -t_1 & p_1 s & & & & & & p_2 \\ \hline & & & & s^2-1 & 1 & & & & \\ \hline & & & & & -t_4 & & & & p_6 \\ \hline & & & & & & 1 & 0 & -2s & 0 \\ & & & & & & 0 & 1 & -2s^2 & s \\ & & & & & & -t_3 & 0 & 0 & p_5 s \\ & & & & & & 0 & -t_5 & 0 & p_7 s^2 \\ & & & & & & 0 & 0 & -t_2 & p_3 + p_4 s \end{array}\right) \end{array}.$$

As claimed, the matrix $\hat{D}$ is a polynomial matrix in s and it agrees with $\bar{D}$ (of Example 2.1) in the diagonal blocks. Also notice the difference between the zero/nonzero structures of $\bar{D}$ and $\hat{D}$. In particular, we can exchange the positions of the two blocks $\{w_1, x_3\}$ and $\{x_4\}$ in $\bar{D}$ without destroying the block-triangular structure if we accordingly exchange the corresponding rows, whereas these two blocks must be arranged in this order in $\hat{D}$ to make it in a block-triangular form. In other words, the square diagonal blocks are partially ordered as $\{w_1, x_3\} \prec \{w_4\}$, $\{x_4\} \prec \{w_4\}$ with respect to the zero/nonzero structure in $\bar{D}$, whereas they are totally ordered as $\{w_1, x_3\} \prec \{x_4\} \prec \{w_4\}$ in $\hat{D}$.

4. Conclusion. We have identified the generic dependency of the Smith normal form on the system parameters $\mathcal{T}$. Then a natural problem is to find an efficient algorithm for computing the Smith normal form, or the invariant polynomials of $D(s) = Q_D(s) + T_D(s)$. This problem has been solved in two special cases. If $T_D(s) = O$, then $D(s)$ is simply a polynomial matrix over $\boldsymbol{K}$, for which Kannan [2] proposed a polynomial-time algorithm. If $Q_D(s)$ satisfies the third condition (A3), which is trivially true in the other extreme case of $Q_D(s) = O$, an efficient (polynomial-time) matroid-theoretic algorithm of Murota [6] is available. A naive idea for attacking the general case would be to substitute numbers from $\boldsymbol{K}$ for the parameters $\mathcal{T}$ and to apply Kannan's algorithm. Then we will often be able to obtain the correct invariant polynomials $e_k(s)$ for $k = 1, \ldots, r-1$, since they are generically independent of $\mathcal{T}$. Elaboration of this idea or others is left for future research.

REFERENCES

[1] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1959.

[2] R. KANNAN, *Solving systems of linear equations over polynomials*, Theoret. Comput. Sci., 39 (1985), pp. 69–88.

[3] K. MUROTA, *Systems Analysis by Graphs and Matroids—Structural Solvability and Controllability*, Algorithms and Combinatorics, Vol. 3, Springer-Verlag, Berlin, 1987.

[4] ———, *Some recent results in combinatorial approaches to dynamical systems*, Linear Algebra Appl., 122/123/124 (1989), pp. 725–759.

[5] ———, *On the irreducibility of layered mixed matrices*, Linear and Multilinear Algebra, 24 (1989), pp. 273–288.

[6] ———, *On the Smith normal form of structured polynomial matrices*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 747–765.

[7] ———, *Mixed matrices—Irreducibility and decomposition*, in Combinatorial and Graph-Theoretical Problems in Linear Algebra, R. A. Brualdi, S. Friedland, and V. Klee, eds., The IMA Volumes in Mathematics and Its Applications, Vol. 50, Springer-Verlag, Berlin, 1993.

[8] K. MUROTA, M. IRI, AND M. NAKAMURA, *Combinatorial canonical form of layered mixed matrices and its application to block-triangularization of systems of equations*, SIAM J. Alg. Discrete Meth., 8 (1987), pp. 123–149.

[9] M. NEWMAN, *Integral Matrices*, Academic Press, New York, 1972.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1112–1123, October 1993

015

ON THE SENSITIVITY OF THE SOLUTION OF NEARLY UNCOUPLED MARKOV CHAINS*

GUODONG ZHANG†

Abstract. This paper deals with the sensitivity of the solution of a nearly uncoupled Markov chain (NUMC). Such a chain arises in various applications where the states to be modeled can be grouped into loosely connected aggregates. The solution of an NUMC is very sensitive to general perturbations. However, in practice the perturbation has a special structure which renders worst case perturbation bounds a large overestimate. In this paper the structure of the perturbation is exploited and it is shown that the solution is very insensitive to a certain class of structured perturbations.

Key words. sensitivity, perturbation theory, nearly uncoupled Markov chain, stochastic matrix

AMS subject classifications. 65F35, 15A51

1. Introduction.

In this paper we will be concerned with the sensitivity of the solution of a nearly uncoupled Markov chain (NUMC) whose transition matrix has the form

$$\mathbf{P} = \mathbf{D} + \mathbf{E} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{E}_{12} & \cdots & \mathbf{E}_{1t} \\ \mathbf{E}_{21} & \mathbf{D}_{22} & \cdots & \mathbf{E}_{2t} \\ \vdots & \vdots & & \vdots \\ \mathbf{E}_{t1} & \mathbf{E}_{t2} & \cdots & \mathbf{D}_{tt} \end{pmatrix}, \tag{1.1}$$

where all the elements of the off-diagonal blocks $\mathbf{E}_{ij}$ are small. We will suppose that $\mathbf{P}$ is irreducible so that it has a unique positive left Perron vector $\boldsymbol{\pi}$ corresponding to the eigenvalue one; that is,

$$\boldsymbol{\pi}^{\mathrm{T}}\mathbf{P} = \boldsymbol{\pi}^{\mathrm{T}}. \tag{1.2}$$

We will normalize $\boldsymbol{\pi}$ so that the sum of its components is one; that is,

$$\boldsymbol{\pi}^{\mathrm{T}}\mathbf{1} = 1,$$

where

$$\mathbf{1}^{\mathrm{T}} = (1,\ 1,\ \ldots,\ 1).$$

We will use the vector $\mathbf{1}$ quite often in the later discussion and only give its size with a subscript (e.g., $\mathbf{1}_n$ for the vector with n elements) explicitly when necessary. Note that $\mathbf{P}$ is stochastic, i.e.,

$$\mathbf{P1} = \mathbf{1}.$$

The sensitivity of the solution for a general Markov chain has been addressed by many authors [11], [10], [6], [5]. Obviously, none of these results can be used to predict the behavior of the solution for an NUMC. The first work addressing the behavior of

* Received by the editors March 4, 1991; accepted for publication (in revised form) December 31, 1991.

† CONVEX Computer Corporation, P.O. Box 833851, Richardson, Texas 75083-3851. This research was done while the author was at the Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland 20742 (zhang@convex.com).

NUMCs appeared in [12]. The subject then received much attention from researchers, see [3], [15], and more recently [16], [13]. Especially, the sensitivity of the solution of an NUMC was analyzed in [16], which gives a deep insight into the transient behavior of the chain. In addition, special computational strategies have been investigated in [18], [2], [8], [19]

The central issue in discussions on the sensitivity of the solution of NUMCs is to assess the difference between the original solution and its perturbed counterpart. Let

$$\tilde{\mathbf{P}} = \mathbf{P} + \mathbf{F}$$

be a perturbed counterpart of $\mathbf{P}$ by $\mathbf{F}$. We assume that $\tilde{\mathbf{P}}$ is still an irreducible stochastic matrix, i.e.,

$$\mathbf{1} = \tilde{\mathbf{P}}\mathbf{1} = \mathbf{P}\mathbf{1} + \mathbf{F}\mathbf{1} = \mathbf{1} + \mathbf{F}\mathbf{1},$$

which implies that $\mathbf{F1} = \mathbf{0}$. Let $\tilde{\boldsymbol{\pi}}$ be the left Perron vector of $\tilde{\mathbf{P}}$. It has been proved in [16] that the solution of NUMCs could be very sensitive under perturbation $\mathbf{F}$ in the sense that the bound on the difference of $\tilde{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ increases in inverse proportion to the size of $\mathbf{E}$. This pessimistic result is anticipated since in the worst case such a perturbation $\mathbf{F}$ could totally contaminate the information carried by $\mathbf{E}$. However, in many practical situations the perturbation itself bears some special structure, see, e.g., [13]. It is the aim of this paper to discuss the sensitivity of $\boldsymbol{\pi}$ to a special class of perturbations. Specifically, we partition $\mathbf{F}$ conformally with $\mathbf{P}$, i.e.,

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} & \cdots & \mathbf{F}_{1t} \\ \mathbf{F}_{21} & \mathbf{F}_{22} & \cdots & \mathbf{F}_{2t} \\ \vdots & \vdots & & \vdots \\ \mathbf{F}_{t1} & \mathbf{F}_{t2} & \cdots & \mathbf{F}_{tt} \end{pmatrix},$$

and assume that

$$\begin{aligned} \|\mathbf{F}_{ii}\| &\le \delta, & i &= 1, \ldots, t, \\ \|\mathbf{F}_{ij}\| &\le \epsilon\delta, & i, j &= 1, \ldots, t, \end{aligned}$$

where ϵ is the upper bound of $\|\mathbf{E}_{ij}\|$, and δ is another small quantity. The symbol $\|\cdot\|$ is used to denote the Euclidean vector norm and the subordinate matrix norm with its specification depending on the context. Since we are mainly concerned with the case when ϵ becomes small, we will say a quantity is of order unity if it is bounded as ϵ tends to zero.

In §2 of this paper we introduce a perturbation expansion that appeared first in [16] and derive an expression suitable for our purpose. In §3 we state two lemmas, which are then used to prove our main theorem in §4. In §5 we discuss the relation of conditions in our analysis and the regularity conditions suggested in [15], [18], [16]. The notation convention is that all matrices and vectors are denoted by bold characters.

2. A perturbation expansion.

Let $\mathbf{P}$ be an $n \times n$ irreducible stochastic matrix and $\boldsymbol{\pi}^{\mathrm{T}}$ be its left Perron vector. To analyze the sensitivity of $\boldsymbol{\pi}^{\mathrm{T}}$ to perturbations, we first introduce a perturbation expansion [16] which gives a relation between $\boldsymbol{\pi}$ and its perturbed counterpart $\tilde{\boldsymbol{\pi}}$. We note that the analysis of this section is valid for a general perturbation $\mathbf{F}$ as long as the perturbed matrix $\tilde{\mathbf{P}}$ is still an irreducible stochastic matrix.

It has been proved in [16] that there exist $n \times (n-1)$ matrices $\mathbf{U}$ and $\mathbf{V}$ such that

$$\tilde{\boldsymbol{\pi}}^{\mathrm{T}} = \boldsymbol{\pi}^{\mathrm{T}} + \boldsymbol{\pi}^{\mathrm{T}}\mathbf{F}\mathbf{U}(\mathbf{I} - \tilde{\mathbf{B}})^{-1}\mathbf{V}^{\mathrm{T}}, \tag{2.1}$$

where

$$\tilde{\mathbf{B}} = \mathbf{V}^{\mathrm{T}}\tilde{\mathbf{P}}\mathbf{U}.$$

The matrices $\mathbf{U}$ and $\mathbf{V}$ can be chosen such that $\|\mathbf{V}\| = 1$, $\|\mathbf{U}\| = \|\mathbf{1}\|\|\boldsymbol{\pi}\|$, and

$$(\mathbf{1},\ \mathbf{U})^{-1} = \begin{pmatrix} \boldsymbol{\pi}^{\mathrm{T}} \\ \mathbf{V}^{\mathrm{T}} \end{pmatrix}.$$

Moreover,

$$\begin{pmatrix} \boldsymbol{\pi}^{\mathrm{T}} \\ \mathbf{V}^{\mathrm{T}} \end{pmatrix} \mathbf{P}(\mathbf{1}\ \ \mathbf{U}) = \begin{pmatrix} \boldsymbol{\pi}^{\mathrm{T}}\mathbf{P}\mathbf{1} & \boldsymbol{\pi}^{\mathrm{T}}\mathbf{P}\mathbf{U} \\ \mathbf{V}^{\mathrm{T}}\mathbf{P}\mathbf{1} & \mathbf{V}^{\mathrm{T}}\mathbf{P}\mathbf{U} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}. \tag{2.2}$$

The perturbation expansion (2.1) shows that it is the term

$$\boldsymbol{\pi}^{\mathrm{T}}\mathbf{F}\mathbf{U}(\mathbf{I} - \tilde{\mathbf{B}})^{-1}\mathbf{V}^{\mathrm{T}}$$

that determines the sensitivity of the solution $\boldsymbol{\pi}$. Since $\|\mathbf{V}\| = 1$, we need only to consider the term

$$\boldsymbol{\Omega} \overset{\text{def}}{=} \mathbf{F}\mathbf{U}(\mathbf{I} - \tilde{\mathbf{B}})^{-1}.$$

Noticing that $\mathbf{V}^{\mathrm{T}}\mathbf{1} = 0$ and

$$\mathbf{U} = (\mathbf{I} - \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}})\mathbf{V},$$

we have

$$\tilde{\mathbf{B}} = \mathbf{V}^{\mathrm{T}}\tilde{\mathbf{P}}\mathbf{V}, \tag{2.3}$$

and

$$\begin{aligned} \boldsymbol{\Omega} &= \mathbf{F}(\mathbf{I} - \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}})\mathbf{V}(\mathbf{I} - \mathbf{V}^{\mathrm{T}}\tilde{\mathbf{P}}\mathbf{V} + \mathbf{V}^{\mathrm{T}}\tilde{\mathbf{P}}\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}\mathbf{V})^{-1} \\ &= \mathbf{F}\mathbf{V}(\mathbf{I} - \tilde{\mathbf{B}})^{-1}. \end{aligned} \tag{2.4}$$

It has been pointed out in [16] that the matrix $\mathbf{V}$ may be chosen to be any matrix with orthonormal columns that span the complement subspace orthogonal to the subspace spanned by $\mathbf{1}$. One of the simplest ways to construct such $\mathbf{V}$ is to take the second to nth columns of the Householder transformation whose first column is parallel to $\mathbf{1}$, which can be explicitly written as

$$\mathbf{V} = \begin{pmatrix} \mathbf{0}^{\mathrm{T}} \\ \mathbf{I}_{n-1} \end{pmatrix} - \begin{pmatrix} \alpha & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & \beta\mathbf{I}_{n-1} \end{pmatrix} \mathbf{1}_n\mathbf{1}_{n-1}^{\mathrm{T}},$$

where

$$\alpha = \frac{1}{\sqrt{n}}, \qquad \beta = \frac{n - \sqrt{n}}{n(n-1)}.$$

By this choice of $\mathbf{V}$, we note that

$$\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}.$$

The inverse of $(\mathbf{I} - \tilde{\mathbf{B}})$ will play an important role in (2.4) when the associated Markov chain becomes nearly uncoupled. When ϵ is small, the matrix $\tilde{\mathbf{P}}$ will have at least $t-1$ eigenvalues close to 1 within the distance of order ϵ. Thus the norm of $(\mathbf{I}-\tilde{\mathbf{B}})^{-1}$ is of order ϵ^{-1}, which indicates that it is unlikely to obtain a tight bound by simply taking the norm in the perturbation expansion (2.1). As a matter of fact, the matrix $\mathbf{V}(\mathbf{I} - \tilde{\mathbf{B}})^{-1}$ turns out to be highly structured, which results in cancellations when it is premultiplied by the matrix $\mathbf{F}$. To exploit this structure, we set

$$\tilde{\boldsymbol{\Omega}} = \mathbf{V}(\mathbf{I} - \tilde{\mathbf{B}})\mathbf{V}^{\mathrm{T}}. \tag{2.5}$$

Since $\mathbf{V}$ has orthogonal columns, we have

$$(\mathbf{I} - \tilde{\mathbf{B}})^{-1} = \mathbf{V}^{\mathrm{T}}\tilde{\boldsymbol{\Omega}}^{\dagger}\mathbf{V},$$

where $\tilde{\boldsymbol{\Omega}}^{\dagger}$ is the Moore–Penrose inverse of $\tilde{\boldsymbol{\Omega}}$. Then,

$$\begin{aligned}\boldsymbol{\Omega} &= \mathbf{F}\mathbf{V}\mathbf{V}^{\mathrm{T}}\tilde{\boldsymbol{\Omega}}^{\dagger}\mathbf{V}\\ &= \mathbf{F}(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\tilde{\boldsymbol{\Omega}}^{\dagger}\mathbf{V}\\ &= \mathbf{F}\tilde{\boldsymbol{\Omega}}^{\dagger}\mathbf{V}.\end{aligned}$$

We are now going to analyze the structure of the Moore–Penrose inverse of $\tilde{\boldsymbol{\Omega}}$ when the associated Markov chain is nearly uncoupled. First, we give two useful lemmas, which will be used to prove our main result.

3. Two lemmas.

For an irreducible stochastic matrix $\mathbf{P}$ associated with a nearly uncoupled Markov chain, we have

$$\mathbf{D}_{ii}\mathbf{1} = \mathbf{1} - \mathbf{e}_i, \qquad i = 1, \ldots, t,$$

where

$$\mathbf{e}_i = \sum_{j\neq i} \mathbf{E}_{ij}\mathbf{1}.$$

It is easy to see that $\mathbf{e}_i \geq \mathbf{0}$ with at least one positive entry. For each $\mathbf{D}_{ii}$, we can construct a stochastic matrix (we drop subscripts for convenience)

$$\bar{\mathbf{D}} = \mathbf{D} + \frac{1}{q}\mathbf{e}\mathbf{1}^{\mathrm{T}}, \tag{3.1}$$

where q is the dimension of $\mathbf{D}$. Since $\mathbf{P}$ is irreducible, it can be seen that $\bar{\mathbf{D}}$ is also irreducible. The following lemma, which was implied in [4], [7], exploits the structure of the inverse of $(\mathbf{I} - \mathbf{D})$.

LEMMA 3.1. *Let $\bar{\boldsymbol{\pi}}$ be the left Perron vector of $\bar{\mathbf{D}}$. Assume*

$$\bar{\boldsymbol{\pi}}^{\mathrm{T}}\mathbf{e} \geq c_1\epsilon, \tag{3.2}$$

for some constant c_1, then the inverse of $\mathbf{I}-\mathbf{D}$ *has the property that*

$$(\mathbf{I}-\mathbf{D})^{-1} = (\mathbf{1},\mathbf{1},\ldots,\mathbf{1})\begin{pmatrix} c^{(1)} & 0 & \cdots & 0 \\ 0 & c^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c^{(q)} \end{pmatrix} + O(1),$$

where the constants $c^{(i)}$, $i=1,2,\ldots,q$, are bounded by

$$|c^{(i)}| \le \frac{c_2}{\epsilon},$$

for some constant c_2.

Proof. From a corollary of the Perron–Frobenius theorem, see, e.g., [1], we know that

$$\rho(\mathbf{D}) < \rho(\bar{\mathbf{D}}) = 1,$$

i.e., the inverse of $(\mathbf{I}-\mathbf{D})$ exists. From

$$\bar{\boldsymbol{\pi}}^{\mathrm{T}}(\mathbf{D} + q^{-1}\mathbf{e}\mathbf{1}^{\mathrm{T}}) = \bar{\boldsymbol{\pi}}^{\mathrm{T}},$$

we have

$$\bar{\boldsymbol{\pi}}^{\mathrm{T}}(\mathbf{I}-\mathbf{D}) = q^{-1}\bar{\boldsymbol{\pi}}^{\mathrm{T}}\mathbf{e}\mathbf{1}^{\mathrm{T}}.$$

By multiplying $(\mathbf{I}-\mathbf{D})^{-1}\mathbf{1}$ on both sides of the above equality, we have

$$\mathbf{1}^{\mathrm{T}}(\mathbf{I}-\mathbf{D})^{-1}\mathbf{1} = \frac{q}{\bar{\boldsymbol{\pi}}^{\mathrm{T}}\mathbf{e}}.$$

Notice that $(\mathbf{I}-\mathbf{D})^{-1}$ is nonnegative. The hypothesis (3.2) ensures that $(\mathbf{I}-\mathbf{D})^{-1}$ is at most the order of ϵ^{-1}, which also indicates that

$$\det(\mathbf{I}-\mathbf{D}) \ge \tilde{c}_1\epsilon, \tag{3.3}$$

for some constant $\tilde{c}_1$ of order unity. Let

$$\mathbf{I}-\mathbf{D} = (\boldsymbol{\xi}_1\ \boldsymbol{\xi}_2\ \cdots\ \boldsymbol{\xi}_q).$$

Then,

$$\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2 + \cdots + \boldsymbol{\xi}_q = O(\epsilon).$$

By the fact of this nearly linear dependency of $\{\boldsymbol{\xi}_i\}$ and the lower bound (3.3), we can obtain the result by using the formula

$$(\mathbf{I}-\mathbf{D})^{-1} = \frac{\operatorname{adj}(\mathbf{I}-\mathbf{D})}{\det(\mathbf{I}-\mathbf{D})},$$

to calculate the inverse. □

We point out that condition (3.2) imposed in the lemma is closely related to the first and second regularity conditions on the NUMCs suggested in [15], [18], [16]. We will return to this point later in the paper. The importance of Lemma 3.1 lies in the fact that the dominant term of the inverse of $(\mathbf{I}-\mathbf{D})$ has constant columns. We will see in the next section that the dominant term of $\tilde{\mathbf{\Omega}}^{\dagger}$ is a partitioned matrix with each portion of its column being a constant vector. The second lemma of this section reveals the cancellation effect of $\mathbf{F}$ when it is used to premultiply a vector with constant segments corresponding to the structure of $\mathbf{F}$.

LEMMA 3.2. *Let column vector* $\boldsymbol{\eta}$ *be partitioned as*

$$\boldsymbol{\eta}^{\mathrm{T}} = (\eta_1\mathbf{1}^{\mathrm{T}},\ \eta_2\mathbf{1}^{\mathrm{T}}, \ldots, \eta_t\mathbf{1}^{\mathrm{T}}),$$

conformed with the block structure of $\mathbf{F}$. *Then,*

$$\|\mathbf{F}\boldsymbol{\eta}\| \le c\epsilon\delta,$$

where c *is a constant of order* $\|\boldsymbol{\eta}\|$.

Proof. The proof is straightforward by noticing that, for $1 \le i \le t$,

$$\mathbf{F}_{i1}\mathbf{1} + \mathbf{F}_{i2}\mathbf{1} + \cdots + \mathbf{F}_{it}\mathbf{1} = \mathbf{0},$$

and

$$\sum_{j\neq i}^{t} \|\mathbf{F}_{ij}\| = O(\epsilon\delta). \qquad \square$$

4. Sensitivity of the solution. Now we come to assess the sensitivity of $\boldsymbol{\pi}$ to the structured perturbation $\mathbf{F}$. We assume that ϵ is small and condition (3.2) is satisfied by all diagonal blocks $\mathbf{D}_{ii}$, $i = 1, \ldots, t$. As pointed out in §2, we need to consider the Moore–Penrose inverse of $\tilde{\mathbf{\Omega}}$.

To this end, we need to work on the singular value decomposition (SVD) of $\tilde{\mathbf{\Omega}}$. From the structure of $\tilde{\mathbf{P}}$, we know that the order of $t-1$ smallest nonzero singular values of $(\mathbf{I} - \tilde{\mathbf{P}})$ is bounded above by ϵ. To see the order of singular values of $\tilde{\mathbf{\Omega}}$, we first note that nonzero singular values of $\tilde{\mathbf{\Omega}}$ are singular values of $(\mathbf{I} - \tilde{\mathbf{B}})$. Noticing that $\tilde{\mathbf{P}}$ itself is an irreducible stochastic matrix with the same characteristic as that of $\mathbf{P}$, we have (cf. (2.2)),

$$\begin{pmatrix} \tilde{\boldsymbol{\pi}}^{\mathrm{T}} \\ \mathbf{V}^{\mathrm{T}} \end{pmatrix} (\mathbf{I} - \tilde{\mathbf{P}})(\mathbf{1}\ \ \tilde{\mathbf{U}}) = \begin{pmatrix} 0 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & \mathbf{I} - \tilde{\mathbf{B}} \end{pmatrix},$$

where $\tilde{\mathbf{U}} = (\mathbf{I} - \mathbf{1}\tilde{\boldsymbol{\pi}}^{\mathrm{T}})\mathbf{V}$. Let

$$\sigma_2 \le \sigma_3 \le \cdots \le \sigma_n$$

and

$$0 = \tilde{\sigma}_1 < \tilde{\sigma}_2 \le \tilde{\sigma}_3 \le \cdots \le \tilde{\sigma}_n$$

be the singular values of $\mathbf{I} - \tilde{\mathbf{B}}$ and $\mathbf{I} - \tilde{\mathbf{P}}$, respectively. Using an inequality concerning the singular values of the product of two matrices (cf. [17, Chap. I, Thm. 4.5]), we obtain the equivalent relation between σ_i and $\tilde{\sigma}_i$, i.e.,

$$c_1\tilde{\sigma}_i \le \sigma_i \le c_2\tilde{\sigma}_i, \quad i = 2, \ldots, n, \tag{4.1}$$

for some constants c_1 and c_2 depending only on the norm of $\mathbf{V}$ and the dimension n. Inequality (4.1) shows that the order of $\sigma_2, \ldots, \sigma_t$ is bounded above by ϵ. From another inequality relating the eigenvalues with the singular values of a matrix (cf. II.4.1.9 in [9]), we have, for eigenvalues λ_i, $i = 2, \ldots, n$, of $\tilde{\mathbf{B}}$,

$$\prod_{i=2}^{n} |1 - \lambda_i| \le \prod_{i=2}^{n} \sigma_i, \tag{4.2}$$

where λ_i, $i = 2, \ldots, n$, are ordered so that

$$0 < |1 - \lambda_2| \le |1 - \lambda_3| \le \cdots \le |1 - \lambda_n|.$$

Under the assumption that $\mathbf{D}_{ii}$, $i = 1, \ldots, t$, satisfy condition (3.2), $\tilde{\mathbf{B}}$ has only $t - 1$ eigenvalues that are near to 1 within the order of ϵ. The inequality (4.2) then suggests that the order of σ_2 cannot be smaller than ϵ if the order of $|1 - \lambda_2|$ is bounded below by ϵ. We now impose the following condition on λ_2:

$$|1 - \lambda_2| \ge c\epsilon, \tag{4.3}$$

for some constant c of order unity. The discussion of the relation of (4.3) to the regularity conditions is deferred to the next section.

The following theorem exploits the structure of the Moore–Penrose inverse of $\tilde{\mathbf{\Omega}}$.

THEOREM 4.1. *Let the* SVD *of* $\tilde{\mathbf{\Omega}}$ *be*

$$\tilde{\mathbf{\Omega}} = \mathbf{S} \begin{pmatrix} 0 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & \mathbf{\Sigma} \end{pmatrix} \mathbf{T}^{\mathrm{T}},$$

where

$$\mathbf{\Sigma} = \mathrm{diag}(\sigma_2,\ \sigma_3, \ldots, \sigma_n).$$

Suppose that $\sigma_2, \ldots, \sigma_t$ *are of order* ϵ, *i.e.,*

$$c_1\epsilon \le \sigma_i \le c_2\epsilon, \quad i = 2, \ldots, t, \tag{4.4}$$

and all diagonal blocks $\tilde{\mathbf{D}}_{ii}$, $i = 1, \ldots, t$, *satisfy the condition* (3.2). *Then,*

$$\tilde{\mathbf{\Omega}}^{\dagger} = (\mathbf{0},\ \sigma_2^{-1}\eta_2, \ldots, \sigma_t^{-1}\eta_t,\ \mathbf{0}, \ldots, \mathbf{0})\mathbf{S}^{\mathrm{T}} + O(1),$$

where

$$\eta_i^{\mathrm{T}} = (\eta_1^{(i)}\mathbf{1}^{\mathrm{T}},\ \eta_2^{(i)}\mathbf{1}^{\mathrm{T}}, \ldots, \eta_t^{(i)}\mathbf{1}^{\mathrm{T}}), \quad i = 2, 3, \ldots, t,$$

for constants $\eta_j^{(i)}$, $j = 1, \ldots, t$, *with the partition conformal with* $\mathbf{P}$.

Proof. From (2.5), we have

$$\begin{aligned} \tilde{\mathbf{\Omega}} &= \mathbf{V}\mathbf{V}^{\mathrm{T}}(\mathbf{I} - \tilde{\mathbf{P}})\mathbf{V}\mathbf{V}^{\mathrm{T}} \\ &= (\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}})(\mathbf{I} - \tilde{\mathbf{P}}). \end{aligned}$$

Let

$$\mathbf{S} = (\mathbf{s}_1,\ \mathbf{s}_2, \ldots, \mathbf{s}_n), \qquad \mathbf{T} = (\mathbf{t}_1,\ \mathbf{t}_2, \ldots, \mathbf{t}_n),$$

then

$$(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}})(\mathbf{I} - \tilde{\mathbf{P}})\mathbf{t}_j = \sigma_j\mathbf{s}_j, \quad j = 2, \ldots, t.$$

Noticing that

$$\left(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)=(n^{-1}\mathbf{1},\ \mathbf{V})\begin{pmatrix}0 & \mathbf{0}^{\mathrm{T}}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\begin{pmatrix}n^{-1}\mathbf{1}^{\mathrm{T}}\\ \mathbf{V}^{\mathrm{T}}\end{pmatrix},$$

we have

$$\mathbf{V}^{\mathrm{T}}[(\mathbf{I}-\tilde{\mathbf{P}})\mathbf{t}_j-\sigma_j\mathbf{s}_j]=0,$$

i.e., $(\mathbf{I}-\tilde{\mathbf{P}})\mathbf{t}_j-\sigma_j\mathbf{s}_j$ is parallel to $\mathbf{1}$. Let

$$(\mathbf{I}-\tilde{\mathbf{P}})\mathbf{t}_j-\sigma_j\mathbf{s}_j=\xi_j\mathbf{1}.$$

Then, by premultiplying with $\tilde{\boldsymbol{\pi}}^{\mathrm{T}}$,

$$\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\mathbf{s}_j=-\frac{\xi_j}{\sigma_j}\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\mathbf{1}.$$

Since $\sigma_j=O(\epsilon)$, $j=2,\ldots,t$, while the left-hand side of the above equality is of order unity, we know that

$$\xi_j=O(\epsilon).$$

Therefore,

$$(\mathbf{I}-\tilde{\mathbf{P}})\mathbf{t}_j=\omega_j\hat{\mathbf{s}}_j,$$

where $\omega_j=O(\epsilon)$, and $\|\hat{\mathbf{s}}_j\|$ is of order unity. Partition $\mathbf{t}_j$ conformally as

$$\mathbf{t}_j=\begin{pmatrix}\mathbf{t}_1^{(j)}\\ \mathbf{t}_2^{(j)}\\ \vdots\\ \mathbf{t}_t^{(j)}\end{pmatrix}.$$

Then, for $1\le k\le t$,

$$\begin{aligned}(\mathbf{I}-\tilde{\mathbf{D}}_{kk})\mathbf{t}_k^{(j)}&=\sum_{p\neq k}\tilde{\mathbf{E}}_{kp}\mathbf{t}_p^{(j)}+(\omega_j\hat{\mathbf{s}}_j)_k\\ &\overset{\mathrm{def}}{=}\omega_k^{(j)}\hat{\mathbf{s}}_k^{(j)},\end{aligned}$$

where $\omega_k^{(j)}=O(\epsilon)$ and $\hat{\mathbf{s}}_k^{(j)}$ is a vector with the norm of order unity. By Lemma 3.1, we have

$$\mathbf{t}_k^{(j)}=\eta_k^{(j)}\mathbf{1}+O(\epsilon),\quad k=1,\ldots,t,$$

where $\eta_k^{(j)}$ is of order unity. Hence,

$$\mathbf{t}_j=\boldsymbol{\eta}_j+O(\epsilon),\quad j=2,\ldots,t.$$

The conclusion of the theorem then comes from the fact that

$$\begin{aligned}\tilde{\boldsymbol{\Omega}}^{\dagger}&=(\mathbf{t}_1,\ \mathbf{t}_2,\ldots,\mathbf{t}_t,\ldots,\mathbf{t}_n)\begin{pmatrix}0 & \mathbf{0}^{\mathrm{T}}\\ \mathbf{0} & \boldsymbol{\Sigma}^{-1}\end{pmatrix}\mathbf{S}^{\mathrm{T}}\\ &=(\mathbf{0},\ \sigma_2^{-1}\mathbf{t}_2,\ldots,\sigma_t^{-1}\mathbf{t}_t,\ \mathbf{0},\ldots,\mathbf{0})\mathbf{S}^{\mathrm{T}}+O(1).\qquad\square\end{aligned}$$

Note that condition (4.3) guarantees that the smallest positive singular value σ_2 is bounded below in the order of ϵ, which justifies the hypothesis (4.4). Recalling that

$$\mathbf{\Omega} = \mathbf{F}\tilde{\mathbf{\Omega}}^{\dagger}\mathbf{V},$$

we have, by using Lemma 3.2,

$$||\mathbf{\Omega}|| \le c\delta,$$

for some constant c.

We now state our perturbation bound of the solution of NUMCs in the following corollary.

COROLLARY 4.2. *Let all diagonal blocks* $\tilde{\mathbf{D}}_{ii}$, $i = 1, \dots, t$, *satisfy the condition* (3.2) *and the second largest eigenvalue of* $\tilde{\mathbf{P}}$ *satisfy the condition* (4.3). *If the structured perturbation* $\mathbf{F}$ *satisfies*

$$\mathbf{F1} = 0,$$

we have

$$\frac{||\tilde{\boldsymbol{\pi}} - \boldsymbol{\pi}||}{||\boldsymbol{\pi}||} \le c\delta, \tag{4.5}$$

for some constant c *of order unity.*

The error bound (4.5) shows that the solution $\boldsymbol{\pi}$ of an NUMC is insensitive to the structured perturbations under some conditions related to the regularity conditions of the associated Markov chain, which is to be discussed in the following section. The most interesting feature of (4.5) is its independence on ϵ, which reveals the fact that if we can obtain information with high relative accuracy both in the aggregates and between the aggregates, we can trust the solution from the measured system.[1]

5. Regularity conditions of an NUMC.

In this section we discuss the relations of conditions (3.2) and (4.3) with three regularity conditions in [15], [18], [16]. To avoid introducing more notations, we choose to state these regularity conditions with an informal description and refer the reader to the original papers [15], [18], [16] for rigorous definitions. We further point out that instead of imposing conditions (3.2) and (4.3) on the perturbed problem as required to prove the theorem, we assume that the conditions are now imposed on the original matrix $\mathbf{P}$ and comment that the qualitative behavior will be the same as ϵ approaches zero.

Regularity condition 1 (the balanced aggregation condition). This condition can be stated by the asymptotic block irreducibility of $\mathbf{P}$. The stochastic matrix $\mathbf{P}$ is asymptotically block irreducible in the sense that there is no block reducible matrix that is within the distance of $o(\epsilon)$ to $\mathbf{P}$ when ϵ approaches zero.

This regularity condition precludes the possibility of an aggregate becoming asymptotically isolated from other aggregates.

To introduce the second regularity condition, we have to describe a behavior of a nearly uncoupled Markov chain which has been observed for a long time [12]; see also [3], [15], [2], [8].

[1] It should also be remarked that the bound (4.5) is valid even when ϵ is not small. In this case, condition (4.3) is sufficient, and condition (3.2) should be discarded.

Let $\mathbf{0}_0$ be an initial state vector of a nearly uncoupled Markov chain. At the timestep τ the transient state vector $\mathbf{0}_\tau$ has a decompsition

$$\mathbf{0}_\tau = \boldsymbol{\pi} + \mathbf{0}_\tau^{(s)} + \mathbf{0}_\tau^{(f)},$$

where $\mathbf{0}_\tau^{(s)}$ and $\mathbf{0}_\tau^{(f)}$ both approach zero as $\tau \to \infty$. But the rate at which $\mathbf{0}_\tau^{(s)}$ approaches zero becomes slower as ϵ goes to zero; while the rate at which $\mathbf{0}_\tau^{(f)}$ tends to zero is virtually independent on ϵ. The transient states $\mathbf{0}_\tau^{(s)}$ and $\mathbf{0}_\tau^{(f)}$ are referred to as slow transient and fast transient, respectively, in the literature.

Regularity condition 2 (the fast transient condition). There is a constant $0 < c < 1$ such that

$$\|\mathbf{0}_\tau^{(f)}\| \le c^\tau,$$

as ϵ approaches zero.

This regularity condition insures that the fast transient approaches zero really fast. The third regularity condition concerns the slow transient.

Regularity condition 3 (the slow transient condition). There is a constant c such that

$$\|\mathbf{0}_\tau^{(s)}\| = O((1 - c\epsilon)^\tau),$$

as $\epsilon \to 0$.

This last regularity condition reflects the requirement that the slow transient cannot be too slow.

The second and third regularity conditions can also be stated in terms of the eigenvalues of $\mathbf{D}_{ii}$ and $(\mathbf{I} - \mathbf{P})$. As $\epsilon \to 0$, the second regularity condition essentially requires that the second largest eigenvalues of $\mathbf{D}_{ii}$, $i = 1, \ldots, t$, are bounded away from 1, and the third regularity condition deters $(t-1)$ smallest positive eigenvalues of $(\mathbf{I} - \mathbf{P})$ from approaching to zero faster than ϵ. The relation between the third regularity condition and our condition (4.3) is now clear since eigenvalues of $(\mathbf{I} - \tilde{\mathbf{B}})$ are asymptotically those of nonzero eigenvalues of $(\mathbf{I} - \mathbf{P})$.

To discuss the relation of condition (3.2) with the regularity conditions 1 and 2, we denote

$$\boldsymbol{\pi}^{\mathrm{T}} = (\nu_1 \boldsymbol{\pi}_1^{\mathrm{T}},\ \nu_2 \boldsymbol{\pi}_2^{\mathrm{T}}, \ldots, \nu_t \boldsymbol{\pi}_t^{\mathrm{T}}),$$

where $\boldsymbol{\pi}_i^{\mathrm{T}} \mathbf{1} = 1$, $i = 1, \ldots, t$, and ν_i, $i = 1, \ldots, t$, are called coupling coefficients. We further let $\hat{\boldsymbol{\pi}}_i$ be the left Perron vector of $\mathbf{D}_{ii}$ with $\hat{\boldsymbol{\pi}}_i^{\mathrm{T}} \mathbf{1} = 1$. It has been shown in [14] that

$$\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i + O(\epsilon).$$

If the second largest eigenvalue of $\mathbf{D}_{ii}$ is bounded away from 1 (regularity condition 2), we know that

$$\bar{\boldsymbol{\pi}}_i = \hat{\boldsymbol{\pi}}_i + O(\epsilon),$$

where $\bar{\boldsymbol{\pi}}_i$ is the left Perron vector of $\bar{\mathbf{D}}_{ii}$ defined in (3.1). Therefore,

$$\bar{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i + O(\epsilon).$$

Condition (3.2) essentially requires that

$$\pi_i^T \mathbf{e}_i \geq c\epsilon, \tag{5.1}$$

for some constant c, which is then guaranteed by the balanced aggregation condition. In fact, if any of $\pi_i^T \mathbf{e}_i$ approaches zero faster than ϵ, the corresponding block row (except the diagonal block) must also approach zero faster than ϵ and the stochastic matrix $\mathbf{P}$ is thus asymptotically reducible. Actually, condition (3.2) is asymptotically equivalent to the regularity conditions 1 and 2. To see this, we note that its asymptotic equivalent (5.1) prevents the second largest eigenvalues of $\mathbf{D}_{ii}$ from being close to 1 in a higher order than ϵ since the determinants of $(\mathbf{I} - \mathbf{D}_{ii})$ are bounded by ϵ^{-1}, as was shown in the proof of Lemma 3.1. The fact that condition (5.1) prevents $\mathbf{P}$ from being asymptotically irreducible is obvious.

Acknowledgments. I am grateful to Pete Stewart for his many encouragements and illuminating suggestions as well as the help in the final refinement of this paper. I would like to thank two anonymous referees for their helpful comments and one of the referees for bringing to my attention [4] and [7]. I also would like to thank Professor Carl Meyer for pointing out an error in the first version of this paper.

REFERENCES

[1] A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] W. L. Cao and W. J. Stewart, *Iterative aggregation/disaggregation techniques for nearly uncoupled markov chains*, J. Assoc. Comput. Mach., 32 (1985), pp. 702–719.

[3] P. J. Courtois, *Decomposability*, Academic Press, New York, 1977.

[4] P. J. Courtois and P. Semal, *Bounds for the positive eigenvectors of nonnegative matrices and for their approximation by decomposition*, J. Assoc. Comput. Mach., 31 (1984), pp. 804–825.

[5] R. E. Funderlic and C. D. Meyer, *Sensitivity of the stationary distribution vector for an ergodic Markov chain*, Linear Algebra Appl., 76 (1986), pp. 1–17.

[6] G. H. Golub and C. D. Meyer, *Using the QR factorization and group inversion to compute, differentiate, and estimate the sensitivity of stationary probabilities for Markov chains*, SIAM J. Algebraic Discrete Methods, 7 (1985), pp. 273–281.

[7] M. Haviv, *An approximation to the stationary distribution of a nearly completely decomposable Markov chain and its error analysis*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 589–594.

[8] ———, *Aggregation/disaggregation methods for computing the stationary distribution of a Markov chain*, SIAM J. Numer. Anal., 24 (1987), pp. 952–966.

[9] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[10] C. D. Meyer, *The condition of a finite Markov chain and perturbation bounds for the limiting probabilities*, SIAM J. Algebraic Discrete Methods, 1 (1980), pp. 273–283.

[11] P. J. Schweitzer, *Perturbation theory and finite Markov chains*, J. Appl. Prob., 5 (1968), pp. 401–413.

[12] H. A. Simon and A. Ando, *Aggregation of variables in dynamic systems*, Econometrica, 29 (1961), pp. 111–138.

[13] G. Stewart and G. Zhang, *On a direct method for the solution of nearly uncoupled Markov chains*, Numer. Math., 59 (1991), pp. 1–11.

[14] G. W. Stewart, *Computable error bounds for aggregated Markov chains*, J. Assoc. Comput. Mach., 30 (1983), pp. 271–285.

[15] ———, *On the structure of nearly uncoupled Markov chains*, in Mathematical Computer Performance & Reliability, G. Iazeolla, P. J. Courtois, and A. Hordijk, eds., Elsevier, Amsterdam, 1984, pp. 287–302.

[16] ———, *Perturbation theory for nearly uncoupled Markov chains*, in Numerical Methods for Markov Chains, W. J. Stewart, ed., North Holland, Amsterdam, 1990, pp. 105–120.

[17] G. W. Stewart and J. Guang Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[18] G. W. STEWART, W. J. STEWART, AND D. F. MCALLISTER, *A two-stage iteration for solving nearly uncoupled Markov chains*, Tech. Report TR-1384, Department of Computer Science, University of Maryland, 1984.

[19] W. J. STEWART AND W. WU, *Iteration and aggregation for the numerical solution of* NCD *Markov chains*, Copper Mountain Conference on Iterative Methods, 1990.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1124–1131, October 1993

016

A COMPONENTWISE PERTURBATION ANALYSIS OF THE QR DECOMPOSITION*

HONGYUAN ZHA†

Abstract. A first-order componentwise perturbation analysis of the QR decomposition is presented. In contrast to the traditional normwise perturbation analysis, the bounds derived are invariant under column scaling of the underlying matrix. As an application, an assessment of the accuracy of the computed QR decomposition using Householder transformations is given.

Key words. QR decomposition, perturbation analysis

AMS subject classifications. 65F05, 65F30, 65G05

1. Introduction.

The QR decomposition is one of the most important decompositions of numerical linear algebra: given a matrix $A \in R^{m\times n}$, which is of full column rank,[1] then there exist an orthonormal matrix Q, and a nonsingular upper triangular matrix R with nonnegative diagonal elements such that

$$A = QR.$$

There are several ways of computing the QR decomposition: using Householder transformations, using Jacobi rotations, using the Gram–Schmidt process. The QR decomposition also has many applications such as computing the singular value decomposition via the Kogbetliantz algorithm [5],[2] computing the orthonormal basis of the column space of a matrix, computing the canonical correlations of matrix pairs [2], and computing the generalized singular value decomposition [5], [7]. For all these applications, the accuracy of the computed Q factor is essential.

The purpose of this paper is to derive some componentwise perturbation bounds for the QR decomposition. There are already several papers dealing with the normwise perturbation analysis of the QR decomposition; among them Stewart's paper is the first to give a thorough analysis [6]. His results were further modified and improved by Sun [11]. An asymptotic expansion of the QR decomposition under row scaling is also given in [8]. A typical result of the normwise perturbation analysis can be roughly stated as follows: Assume that A and $A+E$ are of full column rank, let the QR decomposition of the perturbed matrix $A+E$ be

$$A + E = (Q + W)(R + F), \tag{1}$$

where $Q+W$ is orthogonal and $R+F$ is upper triangular with nonnegative diagonal entries. Then

$$\frac{\|F\|_F}{\|R\|_F} \le \kappa(A)\eta_1 \frac{\|E\|_F}{\|A\|_F}, \tag{2}$$

* Received by the editors February 11, 1992; accepted for publication (in revised form) March 12, 1992. This work is supported by National Science Foundation grant DRC-8412314 and Army contract DAAL-03-90-G-0105.

† Scientific Computing and Computational Mathematics, Stanford University, Stanford, California 94305. Present address, Computer Science Department, Pennsylvania State University, University Park, Pennsylvania 16802 (zha@cs.psu.edu).

[1] There are several modified QR decompositions for the case when A is rank deficient; one of the most widely used is the QR decomposition with column pivoting [2, §5.4.1].

[2] Here the matrix A is first reduced to upper triangular form by the QR decomposition.

and

$$\|W\|_F \le \kappa(A)\eta_2 \frac{\|E\|_F}{\|A\|_F}, \tag{3}$$

where $\kappa(A) = \|A\|_2\|A^+\|_2$ is the spectral condition number of A, and η_1 and η_2 are some constants. For a componentwise perturbation of the form $|E| \le \epsilon|A|$, however, the above bounds may not be very useful. The reason is that they do not reflect the facts that the QR decomposition is invariant under column scaling of A, i.e., the QR decomposition of AD is

$$AD = Q(RD),$$

where D is a positive diagonal matrix; and so is the perturbation matrix E

$$|E|D \le \epsilon|AD|.$$

To illustrate the above assertion, let us look at a concrete example.

Example. Consider the following two matrices:

$$A_1 = \begin{pmatrix} 1 & 1 \\ 0 & 10^{-10} \\ 1 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 10^{10} \\ 0 & 10^{10} \\ 1 & 10^{10} \end{pmatrix}.$$

The spectral condition numbers of A_1 and A_2 are

```
cond_2(A_1)=2.8284e+10,   cond_2(A_2)=2.1213e+10.
```

Computing the QR decomposition of A_1 and A_2 using MATLAB on a Sun workstation with machine precision $O(10^{-16})$, we obtain the following result:

```
Q_1=                                 Q_2=
   -7.0711e-01    1.2539e-06            -7.0711e-01    1.0800e-17
             0   -1.0000e+00                      0   -1.0000e+00
   -7.0711e-01   -1.2539e-06            -7.0711e-01   -1.0800e-17
```

and the errors $e_i = \|Q_i - Q_{\text{exact}}\|_F, \quad (i = 1, 2)$:

```
e_1=1.7732e-06,      e_2=3.0547e-17.
```

Using Wilkinson's classical backward error analysis results [13, Chap. 3], we have

$$\|E\|_F/\|A_i\|_F = O(10^{-16}), \qquad (i = 1, 2),$$

where E is the backward error. With regard to the bound in (3) and the above computed results, the bound for A_1 closely matches the error e_1, while the bound for A_2 is a bad overestimation of e_2. The main reason for this difference is that if we scale out the 10^{10} factor from the last column of A_2, then it is a well-conditioned matrix, and the bound would tightly predict the result for the scaled matrix. Therefore, in conclusion, we need to derive perturbation bounds for the QR decomposition that are invariant under column scaling of the matrix.

As pointed out by one of the referees, the normwise perturbation bounds have the virtue that they do reflect the perturbation when the columns of A are properly scaled, and such scaling is justified since the QR decomposition is invariant under column scaling. For instance, a tighter bound can be obtained in the above example if we first scale out the factor 10^{10} in A_2. This fact is also supported by the following result due to van der Sluis [12]: Let a positive diagonal matrix D_0 be chosen such that all the columns of $A_0 = AD_0$ have unit length, then

$$\kappa(A_0) \leq \sqrt{n} \min\{\kappa(AD) : D = \mathrm{diag}(d_i), d_i > 0\}.$$

This is to say that the condition number of A_0 is at most a factor of $\sqrt{n}$ off the optimal condition number of A under column scaling. To use the normwise bounds properly, we therefore need to first scale the columns of the matrix A. However, it will be clear from the results in §2 that our componentwise perturbation bounds provide the scaling automatically, and hence they render it unnecessary to scale the columns of A beforehand.

The rest of the paper is organized as follows. In §2, we derive our main componentwise perturbation bounds. In §3, we apply the results in §2 to assess the accuracy of the computed QR decomposition using Householder transformations. In §4, we summarize the results and point out some directions for further research.

Notation. If $A = (a_{ij})$, then $|A| := (|a_{ij}|)$. We write $|A| \leq |B|$ if $|a_{ij}| \leq |b_{ij}|$, and it is easy to check that if $A = BC$, then $|A| \leq |B||C|$. A matrix norm $\|\cdot\|$ on $R^{m\times n}$ is monotone, if $|A| \leq |B|$ implies $\|A\| \leq \|B\|$ for all $A, B \in R^{m\times n}$ [4].

2. Perturbation bounds.

In this section, we prove our main results concerning the componentwise perturbation bounds of the QR decomposition.

THEOREM 2.1. *Let A be of full column rank, and $|E| \leq \epsilon G|A|$ such that $A + E$ is also of full column rank, where G is a matrix with nonnegative entries. Also let the QR decomposition of A and $A + E$ be $A = QR$ and $A + E = (Q + W)(R + F)$, respectively. Moreover, for a square nonsingular matrix S, let*

$$\kappa_{BS}(S) = \||S^{-1}| \cdot |S|\|$$

be the Bauer–Skeel condition number [10, p. 128],[3] *and*

$$\eta = \max(\||Q^T||G||Q|\|, \||Q^T||G^T||Q|\|). \tag{4}$$

Then if $\|\cdot\|$ is a consistent monotone norm and

$$\epsilon\eta(\kappa_{BS}(R^{-1}) + \kappa_{BS}(R^T)) < 1, \tag{5}$$

then we have for the Q factor

$$\frac{\|W\|}{\|Q\|} \leq \frac{\epsilon(\|G\| + \eta)(\kappa_{BS}(R^{-1}) + \kappa_{BS}(R^T))}{1 - \epsilon\eta(\kappa_{BS}(R^{-1}) + \kappa_{BS}(R^T))} + O(\epsilon^2),$$

and for the R factor

$$\frac{\|F\|}{\|R\|} \leq \epsilon\eta(\kappa_{BS}(R^{-1}) + \kappa_{BS}(R^T)) + O(\epsilon^2).$$

[3] We notice that $\kappa_{BS}(S)$ also depends on the norm in the definition.

Proof. Since $A+E$ is of full column rank, $(A+E)^T(A+E)$ is symmetric positive definite. Let the Cholesky decomposition of $(A+E)^T(A+E)$ be

$$(A+E)^T(A+E) = (R+F)^T(R+F),$$

where $R+F$ is upper triangular. Then $Q+W := (A+E)(R+F)^{-1}$ is orthonormal. We have

$$A+E = (Q+W)(R+F)$$

the QR decomposition of $A+E$. It is easy to check that we have the following identity:

$$W = -QFR^{-1} + ER^{-1} - WFR^{-1},$$

therefore

$$|W| \leq |Q||FR^{-1}| + |E||R^{-1}| + |W||FR^{-1}|. \tag{6}$$

On the other hand, since

$$R^TF + F^TR + F^TF = A^TE + E^TA + E^TE,$$

it follows that

$$FR^{-1} + R^{-T}F^T + R^{-T}F^TFR^{-1} = Q^TER^{-1} + R^{-T}E^TQ + R^{-T}E^TER^{-1}. \tag{7}$$

Since FR^{-1} is upper triangular,[4] we obtain

$$|FR^{-1}| \leq \epsilon\{|Q^T|G|Q||R||R^{-1}| + (|Q^T|G|Q||R||R^{-1}|)^T\} + O(\epsilon^2).$$

Moreover from

$$|F| = |FR^{-1}R| \leq |FR^{-1}||R|,$$

it follows from the definition of η:

$$\|F\| \leq \epsilon\eta(\||R|\cdot|R^{-1}|\| + \||R^{-T}|\cdot|R^T|\|)\||R|\| + O(\epsilon^2).$$

The above combined with (6) proves the bound for $\|W\|$. □

Remark 1. Most of the familiar norms we use are monotone norms, for example, $\|\cdot\|_1$, $\|\cdot\|_\infty$, and $\|\cdot\|_F$. But the spectral norm $\|\cdot\|_2$ is not monotone. One way of getting around this is to modify the spectral norm to obtain $\|A\| := \||A|\|_2$, which is monotone, and the above results can also be cast in terms of this norm.

Remark 2. The bounds in Theorem 2.1 are first-order bounds, and the question when the second-order terms are negligible naturally arises. Recently Stewart gave a condition on the perturbation under which one can derive a bound on the second-order terms [9]. In our context, the condition reduces to the requirement that

$$\epsilon\eta(\kappa_{BS}(R^{-1}) + \kappa_{BS}(R^T)) < \frac{1}{4}.$$

We notice that this condition is stronger than that in (5); this is to guarantee that the second-order term will not blow up [9]. However, we will not elaborate on the details of the proof here.

[4] This trick is due to Stewart [6].

COROLLARY 2.2. *Let the ith column of R and F be r_i and f_i, and let the first i columns of Q and the ith leading submatrix of R be Q_i and R_i, $(i = 1, \cdots n)$, then we have for the $\|\cdot\|_2$ norm*

$$\|f_i\|_2 \leq 2\epsilon n\|r_i\|_2\|G\|_2\big\||R_i|\cdot|R_i^{-1}|\big\|_2 + O(\epsilon^2),$$

and for the $\|\cdot\|_\infty$ norm

$$\|f_i\|_\infty \leq \epsilon n\|r_i\|_\infty \max(\|G\|_\infty, \|G\|_1)\big(\big\||R_i|\cdot|R_i^{-1}|\big\|_\infty + \big\||R_i|\cdot|R_i^{-1}|\big\|_1\big) + O(\epsilon^2),$$

and for the $\|\cdot\|_1$ norm

$$\|f_i\|_1 \leq \epsilon n\|r_i\|_1 \max(\|G\|_\infty, \|G\|_1)\big(\big\||R_i|\cdot|R_i^{-1}|\big\|_\infty + \big\||R_i|\cdot|R_i^{-1}|\big\|_1\big) + O(\epsilon^2).$$

Proof. If A is replaced by AD, where D is a positive diagonal matrix, the right-hand sides of the inequalities in Theorem 2.1 remain the same. Let the ith diagonal element of D be one, and let the other diagonal elements tend to zero. It can be readily checked from (7) that this will not blow up the second-order term. Then the above bounds follow easily after some manipulation. □

The above bounds relate each individual column of F to the corresponding column of R. They are better bounds than those in Theorem 2.1, since a bad scaling of a column of A will not affect the bounds for the other columns.

From the above analysis, we can consider

$$\kappa_S(A) := (\kappa_{BS}(R^{-1}) + \kappa_{BS}(R^T))/2$$

as the componentwise condition number for the QR decomposition. It is easy to see that this condition number is invariant under any column scaling of A; more precisely

$$\kappa_S(AD) = \kappa_S(A),$$

for any positive diagonal matrix D.

Remark 3. Assuming that $A \in R^{m\times n}$, it is easy to check that

$$\big\||Q|\big\|_2 \leq \big\||Q|\big\|_F = \|Q\|_F = \sqrt{n}, \quad \big\||Q|\big\|_\infty \leq \sqrt{n}, \quad \big\||Q|\big\|_1 \leq \sqrt{m}.$$

Remark 4. For a matrix norm $\|\cdot\|$, define

$$\kappa_B(A) := \min\{\|AD\|\cdot\|(AD)^+\| : D = \mathrm{diag}(d_i), d_i > 0\},$$

which is the best condition number of A under column scaling. For the spectral norm, it is easy to show the following inequality

$$\begin{aligned}\kappa_S(A) &= \big\||Q^TQRD|\cdot|(RD)^{-1}Q^TQ|\big\|_2 \\ &\leq \big\||Q^T|\big\|_2\big\||Q|\big\|_2\big\||AD|\big\|_2\big\||(AD)^+|\big\|_2\end{aligned}$$

for any positive diagonal D, hence

$$\kappa_S(A) \leq n^2\kappa_B(A).$$

Similar results for the $\|\cdot\|_1$ and $\|\cdot\|_\infty$ can also be proved.

3. The Householder triangularization. As mentioned in the preface of the book *Matrix Perturbation Theory*, one of the principal purposes of deriving perturbation bounds is to combine them with the backward error analysis to assess the effects of the rounding errors on the computed results [10, p. xiii]. In this section, we apply the results in the previous section to derive some bounds for the QR decomposition computed by the Householder transformations. To this end we need a backward componentwise error analysis of the Householder triangularization method. This analysis was recently carried out by Higham [3]. In the following we use the notation as in the appendix of the paper [3]. Let u be the machine precision, and denote

$$\gamma_n = \frac{nu}{1-nu}.$$

We cite the following result of Higham [3].[5]

LEMMA 3.1. *Consider applying a sequence of Householder transformations P_k to the matrix $A \in R^{m\times n}$,*

$$A_{k+1} = P_k A_k, \quad k = 1, \ldots, n, \tag{8}$$

with $A_1 = A$. Let the computed version of P_k and A_k be $\hat{P}_k$ and $\hat{A}_k$, respectively. Let P_k' be the matrix corresponding to the exact application of the kth step of (8) *to $\hat{A}_k$ defined above.*[6] *Then*

$$\hat{A}_{n+1} = (P_n' \cdots P_1')^T A + \Delta,$$

where $|\Delta| \le \mu_{m,n} G|A|$, and $\mu_{m,n} = \gamma_{am+bn+c}$, with a, b, c small integers. The matrix G satisfies

$$\|G\|_2 \le 2mn(1+\theta_{m,n}),$$

where

$$\theta_{m,n} = \frac{2\mu_{m,n}\sqrt{m}(n-1)}{1-2\mu_{m,n}\sqrt{m}(n-1)}.$$

To derive our bounds, we also need a bound for the difference

$$K = \hat{Q} - P_n' \cdots P_1',$$

where $\hat{Q} = fl(\hat{P}_n \cdots \hat{P}_1)$ is the computed product. The bound can be derived from the general analysis given by Wilkinson [13, p. 161], and is presented in the following lemma.

LEMMA 3.2. *With the assumption the same as in the above lemma, we have*

$$\|\hat{Q} - P_n' \cdots P_1'\|_F \le 2\sqrt{n}\mu_n \left(1 + \frac{2\mu_n(n-2)}{1-2\mu_n(n-2)}\right), \tag{9}$$

where $\mu_n = \gamma_{a'n+b'}$, and a' and b' are small integers.

Combining these two results with the bounds in Theorem 2.1, we have the following conclusion, which we write in the form of a theorem.

[5] This result appears as Lemma A.7 in an extended version of [3] with the same title: Numerical Analysis Report No. 182, University of Manchester, Manchester, U.K., 1990.

[6] For a more elaborate description, see [13, p. 124].

THEOREM 3.3. *Let the assumptions be the same as above, and furthermore assume that*

$$8mn^2\mu_{m,n}(1+\theta_{m,n})\left\||R|\cdot|R^{-1}|\right\|_2 < 1. \tag{10}$$

Then the computed Q factor $\hat{Q}$ satisfies[7]

$$\|\hat{Q}-Q\|_F \lesssim 8mn^{3/2}(1+n)(1+\theta_{m,n})\mu_{m,n}\left\||R|\cdot|R^{-1}|\right\|_2 + O(\mu^2_{m,n}).$$

Proof. We use the trivial identity

$$\hat{Q}-Q = (\hat{Q}-P'_n\cdots P'_1) + (P'_n\cdots P'_1 - Q),$$

and ignore the first parenthesized term, since by (9) it is negligible compared to the bound to be proven. To get the bound for the second term in the above, we use Theorem 2.1 with the spectral norm (see also the comments in Remark 1). We notice that η as defined in (4) has the upper bound

$$\eta \le \|Q\|_F^2\|G\|_2 \le 2mn^2(1+\theta_{m,n}),$$

therefore condition (10) implies that the denominator in the bound for the Q factor is at least one-half. Then the above approximate bound can be obtained by the bounds of Theorem 2.1. □

The results of the example in the introduction section can be well explained by the above bound. In fact, for the spectral norm, we have

$$\kappa_S(A_1) = \texttt{2.8284e+10}, \quad \kappa_S(A_2) = \texttt{3.1463e+00}.$$

4. Concluding remarks. In this paper, we derived some first-order componentwise perturbation bounds for the QR decomposition. Those bounds are invariant under the column scaling of the matrix. Combining those bounds with the result of a componentwise backward error analysis, we also derived bounds for the computed QR factors using the Householder transformations. There are still some problems that deserve further investigation; for example, how to directly estimate the condition number $\kappa_S(A)$, for the $\|\cdot\|_\infty$ or $\|\cdot\|_1$ norm[8] and how to derive bounds for the computed QR decomposition using the Gram–Schmidt process. Another interesting problem is to derive bounds for the second-order terms in Theorem 2.1 in terms of the condition number $\kappa_S(A)$. These problems will be addressed in future research.

Acknowledgment. The author thanks Dr. N. J. Higham and the referees for many helpful comments and suggestions that considerably improved the presentation.

REFERENCES

[1] M. ARIOLI, J. W. DEMMEL, AND I. S. DUFF, *Solving sparse linear systems with sparse backward error,* SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.

[2] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations,* The Johns Hopkins University Press, 2nd ed., Baltimore, MD, 1989.

[7] The comments in Remark 2 about the negligibility of the second-order term still apply here.

[8] A comment from N. J. Higham is relevant here: Since $\left\||R|\cdot|R^{-1}|\right\|_\infty \le n\left\||R^{-T}|\cdot|R^T|\right\|_\infty$, a trick in [1] can be used to estimate the right-hand-side quantity, which in turn provides a method for estimating the above-mentioned condition number.

[3] N. J. HIGHAM, *Iterative refinement enhances the stability of QR decomposition methods for solving linear equations,* BIT, 31 (1991), pp. 447–468.

[4] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis,* Cambridge University Press, London, 1985.

[5] C. C. PAIGE, *Computing the generalized singular value decomposition*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1126–1146.

[6] G. W. STEWART, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.

[7] ———, *Computing the CS-decomposition of a partitioned orthonormal matrix,* Numer. Math., 40 (1982), pp. 297–306.

[8] ———, *On the asymptotic behavior of scaled singular value and QR decompositions,* Math. Comp., 43 (1984), pp. 483–489.

[9] ———, *On the perturbation of LU, Cholesky, and QR factorizations,* Tech. Rep., TR-2848, Dept. of Computer Science, University of Maryland, Baltimore, MD, 1992.

[10] G. W. STEWART AND J. G. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[11] J. G. SUN, *Perturbation bounds for the Cholesky and QR factorizations*, BIT, 31 (1991), pp. 341–352.

[12] A. VAN DER SLUIS, *Condition numbers and equilibration of matrices,* Numer. Math., 14 (1969), pp. 14–23.

[13] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, England, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1132–1140, October 1993

017

TOTAL POSITIVITY, QR FACTORIZATION, AND NEVILLE ELIMINATION*

M. GASCA[†] AND J. M. PEÑA[‡]

Abstract. A well-known characterization of nonsingular totally positive matrices is improved: Only the sign of minors with consecutive initial rows or consecutive initial columns has to be checked. On the other hand, a new characterization of such matrices by their QR factorization is obtained. As in other recent papers of the authors, Neville elimination plays an essential role.

Key words. total positivity, matrix factorization

AMS subject classifications. 65F30, 15A23, 15A48

1. Introduction and notations. We recall that a matrix A is said to be totally positive (respectively, strictly totally positive) if and only if all its minors are nonnegative (respectively, positive). Totally positive (respectively, strictly totally positive) matrices will be referred to as TP (respectively, STP) matrices.

Following usual notations (see [1], [5]), given k, $n \in \mathbb{N}$, $k \leq n$, $Q_{k,n}$ will denote the set of strictly increasing sequences of k natural numbers less than or equal to n:

$$\alpha = (\alpha_i)_{1 \leq i \leq n} \in Q_{k,n} \quad \text{if } (1 \leq)\, \alpha_1 < \alpha_2 < \cdots < \alpha_k \,(\leq n).$$

The dispersion $d(\alpha)$ of α is defined by

$$d(\alpha) := \sum_{i=1}^{k-1} (\alpha_{i+1} - \alpha_i - 1) = \alpha_k - \alpha_1 - (k-1),$$

with the convention $d(\alpha) = 0$ for $\alpha \in Q_{1,n}$. In general, $d(\alpha) = 0$ means that α consists of k consecutive integers.

Let n, m, k, l be natural numbers with $k \leq n$, $l \leq m$. For an $n \times m$ real matrix A and $\alpha \in Q_{k,n}$, $\beta \in Q_{l,m}$ we denote by $A[\alpha|\beta]$ the $k \times l$ submatrix of A containing rows numbered by α and columns numbered by β. When $\alpha = \beta$, $A[\alpha|\alpha]$ will be denoted by $A[\alpha]$.

In [5, Thms. 3.2 and 4.1] we give one characterization of nonsingular TP matrices and another for STP matrices in terms of the sign of some of their minors. Our characterization of TP matrices improved some previous ones, which used a greater number of minors (see [1], [3]): A nonsingular matrix A of order n is TP if and only if it satisfies, simultaneously, the following conditions:

$$\begin{aligned} &\det A[\alpha|\beta] \geq 0 \quad \forall \alpha, \beta \in Q_{k,n} \\ &\qquad \text{such that} \quad d(\alpha) = 0 \quad \text{and either} \quad \alpha_1 = 1 \quad \text{or} \quad \alpha_l > \beta_l \quad \forall l \leq k, \end{aligned} \tag{1.1}$$

*Received by the editors September 23, 1991; accepted for publication March 2, 1992.

[†] Departamento de Matemática Aplicada, Universidad de Zaragoza, Zaragoza, Spain (GASCA@CC.UNIZAR.ES). The work of this author was partially supported by Comision Interministerial de Ciencia y Tecnologia research grant PS87-0060.

[‡] Departamento de Matemática Aplicada, Universidad de Zaragoza, Zaragoza, Spain (JMPENA@CC.UNIZAR.ES). The work of this author was partially supported by Dirección General de Ciencia y Tecnologia research grant PS88-0085.

$$\det A[1,2,\dots,k] > 0 \quad \forall k \in \{1,2,\dots,n\}. \tag{1.2}$$

Compare this characterization with [3], where it is said that A is TP if and only if

$$\det A[\alpha|\beta] \geq 0 \quad \forall \alpha, \beta \in Q_{k,n} \quad \text{such that} \quad d(\alpha) = 0.$$

The analogous characterization of STP matrices in [5, Thm. 4.1], particularized for square matrices, states that a matrix A of order n is STP if and only if it satisfies simultaneously, for each $k \in \{1,2,\dots,n\}$, the following conditions:

$$\det A[\alpha|1,2,\dots,k] > 0 \quad \forall \alpha \in Q_{k,n} \quad \text{with} \quad d(\alpha) = 0, \tag{1.3}$$

$$\det A[1,2,\dots,k|\beta] > 0 \quad \forall \beta \in Q_{k,n} \quad \text{with} \quad d(\beta) = 0. \tag{1.4}$$

The asymmetry of conditions (1.1), (1.2), which is contrary to the symmetry of (1.3), (1.4), led us to search for the new simplification of (1.1) given in Theorem 3.1 of this paper.

Another well-known characterization of TP matrices ([3, Thm. 1.1]) is in terms of their LU factorization: An $n \times n$ matrix A is TP if and only if A has an LU factorization such that L and U are TP. Here, as usual, L (respectively, U) is a lower (respectively, upper) triangular matrix.

However, we have not seen in the literature any result on the QR factorization of TP matrices, that is, their factorization as a product of an orthogonal matrix Q by an upper-triangular matrix R. Because of the interest of that factorization in numerical analysis, we think that it is worthwhile to study it. In §4 we obtain a characterization of nonsingular TP matrices in those terms, but in order to get it we have to define some classes of matrices that contain the class of nonsingular TP matrices. In §5 we characterize one of these classes, the so-called γ-matrices, from an algorithmic point of view: the Neville elimination (NE).

Finally, in §6 we see that an oscillatory matrix A (that is, a TP matrix such that A^m is STP for some positive integer m) can be transformed by similarity into diagonal form D : $P^{-1}AP = D$ with P a γ-matrix. In [7] it is proved that if A is a nonsingular TP matrix (the condition of nonsingularity is removed in [3]), then there exist a TP matrix S and a tridiagonal TP matrix T such that $SAS^{-1} = T$.

2. Neville elimination. The essence of NE is to make zeros in a column of a matrix by adding to each row a multiple of the previous one. Eventual reorderings of the rows of the matrix may be necessary. This process, although not new, is described in a precise way in [5], where some other references on the subject are given.

Because of the frequent reference to some concepts defined in [5], we recall them here, although the concepts are particularized to nonsingular matrices, which are our present subjects. Let $A = (a_{ij})_{1\leq i,j\leq n}$ be a nonsingular matrix, and let $\tilde{A}_1 = \left(\tilde{a}^1_{ij}\right)_{1\leq i,j\leq n}$ be such that $\tilde{a}^1_{ij} := a_{ij}$. If there are zeros in the first column of $\tilde{A}_1$, we carry the corresponding rows down to the bottom in such a way that the relative order among them is the same as in $\tilde{A}_1$. We denote this new matrix by $A_1 = \left(a^1_{ij}\right)_{1\leq i,j\leq n}$. If we do not need row exchanges, then $A_1 := \tilde{A}_1$.

The method consists of constructing a finite sequence of matrices A_k such that the submatrix formed by the $k-1$ initial columns of A_k is an upper-triangular matrix. If $A_t = \left(a^t_{ij}\right)_{1\leq i,j\leq n}$, then we introduce zeros in its tth column below the main diagonal, thus forming

$$\tilde{A}_{t+1} = \left(\tilde{a}^{t+1}_{ij}\right)_{1\leq i,j\leq n}.$$

For any j such that $1 \le j \le n$ we have

$$(2.1) \qquad \begin{cases} \tilde{a}_{ij}^{t+1} := a_{ij}^{t}, & i = 1, 2, \dots, t, \\ \tilde{a}_{ij}^{t+1} := a_{ij}^{t} - (a_{it}^{t}/a_{i-1,t}^{t})a_{i-1,j}^{t} & \text{if } a_{i-1,t}^{t} \neq 0, \quad t < i \le n, \\ \tilde{a}_{ij}^{t+1} := a_{ij}^{t} & \text{if } a_{i-1,t}^{t} = 0, \quad t < i \le n. \end{cases}$$

If $\tilde{A}_{t+1}$ has zeros in the $(t+1)$th column in the row $t+1$ or below it, we will carry these rows down in a manner similar to that we have used with $\tilde{A}_1$. The matrix obtained in this way will be denoted by

$$A_{t+1} = \left(a_{ij}^{t+1}\right)_{1 \le i,j \le n},$$

being $A_{t+1} := \tilde{A}_{t+1}$ if there are no row exchanges in $\tilde{A}_{t+1}$. After at most n steps we get an upper-triangular matrix $A_n = U$.

The element

$$(2.2) \qquad p_{ij} := a_{ij}^{j}, \qquad 1 \le i, j \le n,$$

(see its function in (2.1)) is called the (i,j) pivot of the NE of A, and the (i,j) multiplier is defined by

$$(2.3) \qquad \begin{cases} a_{ij}^{j}/a_{i-1,j}^{j} & \text{if } a_{i-1,j}^{j} \neq 0, \\ 0 & \text{if } a_{i-1,j}^{j} = 0 \left(\Longrightarrow a_{ij}^{j} = 0\right). \end{cases}$$

The complete Neville elimination (CNE) of A consists of carrying out the NE of A until we arrive at U and, afterwards, proceeding with the NE of U^T (the transpose of U) until we get a diagonal matrix. Obviously, the NE of U^T is equivalent to the NE of U by columns.

In [5, Thm. 4.1 and Cor. 5.5] we see that a nonsingular matrix A is TP (respectively, STP) if and only if there are no row or column exchanges in the CNE of A and the pivots are nonnegative (respectively, positive).

3. A simplified characterization of nonsingular TP matrices by minors.

The following theorem provides a characterization of nonsingular TP matrices that is analogous to that of STP matrices given by (1.3) and (1.4).

THEOREM 3.1. *Let A be a nonsingular matrix of order n. Then A is* TP *if and only if it satisfies simultaneously, for each $k \in \{1, 2, \dots, n\}$, the following conditions*:

$$(3.1) \qquad \det A[\alpha|1, 2, \dots, k] \ge 0 \quad \forall \alpha \in Q_{k,n},$$

$$(3.2) \qquad \det A[1, 2, \dots, k|\beta] \ge 0 \quad \forall \beta \in Q_{k,n},$$

$$(3.3) \qquad \det A[1, 2, \dots, k] > 0.$$

Proof. Let A be nonsingular and TP. Inequalities (3.1) and (3.2) are direct consequences of the total positivity of A and (3.3) (see [1, Cor. 3.8]) of the nonsingularity of a TP matrix.

Conversely, from (3.3) A can be decomposed by Gaussian elimination in the form $A = L_1 D_1 U_1$ with L_1 (respectively, U_1) a lower- (respectively, upper-) triangular, unit-diagonal matrix. By unit diagonal we mean a matrix whose diagonal entries are 1.

From the Cauchy–Binet identity (see, for example, [1, Eq. (1.23)]) one has $\forall \alpha \in Q_{k,n}$, $k = 1, 2, \dots, n$,

$$\det(L_1 D_1)[\alpha|1, 2, \dots, k] = \det A[\alpha|1, 2, \dots, k], \tag{3.4}$$

and by (3.1) and [3, Thm. 1.4], $L_1 D_1$ is TP; therefore, D_1 and L_1 are TP.

Since by the same reasons as for (3.4) we have $\forall \beta \in Q_{k,n}$ and $\forall k \in \{1, 2, \dots, n\}$

$$\det(D_1 U_1)[1, 2, \dots, k|\beta] = \det A[1, 2, \dots, k|\beta], \tag{3.5}$$

we infer from (3.2) that $D_1 U_1$ (and U_1) is TP. Then A is TP as a product of TP matrices [1, Thm. 3.1]. □

4. A characterization of nonsingular TP matrices by their *QR* factorization. Let A be an $n \times n$ lower- (respectively, upper-) triangular matrix. Following [2], the minors $A[\alpha|\beta]$ with $\alpha_k \geq \beta_k$ (respectively, $\alpha_k \leq \beta_k$ $\forall k$) are called nontrivial minors of A because all other minors are equal to zero.

In [2] the matrix A is called ΔSTP if and only if the nontrivial minors of A are all positive (see also [6, §9]).

We are going to define some special classes of matrices. In this section, L (respectively, U) represents a lower- (respectively, upper-) triangular, unit-diagonal matrix, and D represents a diagonal matrix.

DEFINITION 4.1. A nonsingular matrix A is said to be lowerly TP (respectively, lowerly STP) if and only if it can be decomposed in the form $A = LDU$ and LD is TP (respectively, ΔSTP).

Observe that since L is unit diagonal, the total positivity of LD implies that L and D are TP.

The following propositions characterize lowerly TP and STP matrices.

PROPOSITION 4.2. *A nonsingular $n \times n$ matrix A is lowerly* TP *if and only if it satisfies for each $k \in \{1, 2, \dots, n\}$ conditions* (3.1) *and* (3.3).

Proof. As in Theorem 3.1, we prove that (3.1) and (3.3) imply that A can be decomposed $A = LDU$ and that LD is TP. Then A is lowerly TP.

Conversely, by definition, if A is lowerly TP, it can be factorized $A = LDU$ with D nonsingular, and then it satisfies

$$\det A[1, 2, \dots, k] \neq 0, \qquad k = 1, 2, \dots, n.$$

As in (3.4), we deduce that these minors are positive and also satisfy (3.1). □

Remark 4.3. If we define upperly TP matrices similarly to lowerly TP matrices, as nonsingular matrices that can be decomposed in the form LDU with DU totally positive, we obviously have that a nonsingular matrix is TP if and only if it is lowerly and upperly TP.

PROPOSITION 4.4. *A nonsingular $n \times n$ matrix A is lowerly* STP *if and only if it satisfies for each $k \in \{1, 2, \dots, n\}$*

$$\det A[\alpha|1, 2, \dots, k] > 0 \quad \forall \alpha \in Q_{k,n} \text{ with } d(\alpha) = 0. \tag{4.1}$$

The proof is analogous to that of Proposition 4.2, where the characterization of ΔSTP matrices given by [2, Thm. 3.1] is taken into account.

DEFINITION 4.5. A nonsingular matrix A is said to be a γ-matrix (respectively, a strict γ-matrix) if it is lowerly TP (respectively, lowerly STP) and in the factorization $A = LDU$, U^{-1} is TP (respectively, ΔSTP).

PROPOSITION 4.6. *If A and $(A^T)^{-1}$ are lowerly* TP (*respectively, lowerly* STP), *then A is a γ-matrix* (*respectively, strict γ-matrix*).

Proof. Let us consider the nonstrict case. We have to see that under the conditions of the proposition, U^{-1} is totally positive.

From the factorization $A = LDU$ we get

$$(U^T)^{-1} = DL^T(A^T)^{-1}. \tag{4.2}$$

DL^T is TP because A is nonsingular, lowerly TP. Then, as we recalled in §2, it is possible to perform the Neville elimination of DL^T without changes of columns, and the pivots are nonnegative.

So we can postmultiply DL^T by certain upper-bidiagonal elementary matrices with a negative off-diagonal entry, that is, matrices of the form

$$\begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & \alpha & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \tag{4.3}$$

with α negative, to get a diagonal matrix. Then DL^T can be written in (4.2) as a product of a diagonal matrix with a finite sequence of inverses of matrices (4.3), which are also of the type (4.3), but with positive off-diagonal entry α.

Therefore, in (4.2) $(U^T)^{-1}$ is obtained from the lowerly TP matrix $(A^T)^{-1}$ by a sequence of transformations that consists of multiplying a row by a positive number and/or adding to a row the next row multiplied by a positive number. Since by Proposition 4.2 all minors of $(A^T)^{-1}$ with initial consecutive columns are nonnegative, all minors of $(U^T)^{-1}$ are also nonnegative. Then $(U^T)^{-1}$ is totally positive [3, Thm. 1.4], as is U^{-1}, and A is a γ-matrix.

The case lowerly STP is proved analogously, by replacing Proposition 4.2 by Proposition 4.6 and by replacing Theorem 1.4 of [3] by Theorem 3.1 of [2]. □

The following theorem gives a characterization of nonsingular TP or STP matrices by their QR factorization.

THEOREM 4.7. *Let A be a nonsingular matrix. Then A is* TP (*respectively,* STP) *if and only if there exist two orthogonal γ-matrices* (*respectively, strict γ-matrices*) Q_1, Q_2, *and two nonsingular, upper-triangular TP* (*respectively,* ΔSTP) *matrices* R_1, R_2 *such that*

$$A = Q_1R_1, \qquad A^T = Q_2R_2. \tag{4.4}$$

Proof. Let A be a nonsingular TP matrix. The Gram–Schmidt method, which permits orthogonalization of the columns of A, is described by

$$AS = H, \tag{4.5}$$

where S is an upper-triangular, unit-diagonal matrix and H is a matrix with orthogonal columns. Then there exists a diagonal matrix D with positive diagonal entries such that $HD = Q_1$ is an orthogonal matrix.

Since A satisfies (3.1) and (3.3), from (4.5) and the Cauchy–Binet formula we deduce that those inequalities hold for H and Q_1 also. Then by Proposition 4.6, Q_1 is a γ-matrix because it is lowerly TP and $(Q_1^T)^{-1} = Q_1$.

On the other hand, from (4.5) we get

$$A = Q_1 D^{-1} S^{-1}, \tag{4.6}$$

and the matrix $R_1 = D^{-1}S^{-1}$ is upper triangular with positive diagonal. Then $A = Q_1 R_1$ and

$$A^T A = R_1^T R_1. \tag{4.7}$$

$R_1^T R_1$ can be computed from R_1 by multiplying each row by a positive number and adding a linear combination of the previous rows. Therefore, minors of $R_1^T R_1$ with consecutive initial rows have the same sign as the corresponding minors of R_1. Since $R_1^T R_1$ is totally positive by (4.7) (as a product of TP matrices), then those minors of $R_1^T R_1$ and R_1 are nonnegative, and by [3, Thm. 1.4], R_1 is TP.

Conversely, assume that A and A^T can be expressed in the form

$$A = Q_1 R_1, \qquad A^T = Q_2 R_2,$$

with Q_1, R_1, Q_2, R_2 as in the theorem. Since by Proposition 4.2, Q_1 satisfies (3.1) and (3.3), the computation of A from Q_1 shows that A satisfies (3.1) and (3.3) also, and the same happens with A^T. By Theorem 3.1, A is TP.

The case of STP matrices is proved in the same way, with adequate replacements as follows. At the end of the first part of the proof, Theorem 1.4 of [3] is replaced by Theorem 3.1 of [2]. In the converse, Proposition 4.2 and Theorem 3.1 of this paper are replaced, respectively, by Proposition 4.4 of this paper and Theorem 4.1 of [5]. □

Remark 4.8. Since the QR factorization is essentially unique, a decomposition (4.4) can be found by any method; then one must choose the appropriate signs.

5. Characterization of strict γ-matrices by NE.

First we observe some facts about NE when it is possible to carry it out without changes of rows. In that case, the NE process for a nonsingular matrix A can be matricially described in the form

$$\begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & 0 & 1 & \\ & & & -m_{n,n-1} & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ & -m_{32} & 1 & & \\ & & \ddots & \ddots & \\ & & & -m_{n2} & 1 \end{bmatrix} \begin{bmatrix} 1 & & & & \\ -m_{21} & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & -m_{n1} & 1 \end{bmatrix} A = U, \tag{5.1}$$

where m_{ij} is the (i,j) multiplier (see §2).

Conversely, if for a nonsingular matrix A one has

$$\begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & 0 & 1 & \\ & & & \alpha_{n,n-1} & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ & \alpha_{32} & 1 & & \\ & & \ddots & \ddots & \\ & & & \alpha_{n2} & 1 \end{bmatrix} \begin{bmatrix} 1 & & & & \\ \alpha_{21} & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \alpha_{n1} & 1 \end{bmatrix} A = U \tag{5.2}$$

with U upper triangular and $\alpha_{ij} \neq 0 \; \forall (i,j)$, then the NE of A can be performed without changes of rows and the multipliers are

$$m_{ij} = -\alpha_{ij}, \qquad i = n, \dots, j+1, \quad j = 1, 2, \dots, n-1.$$

On the basis of this observation, we shall prove the following lemma.

LEMMA 5.1. *Let L be a lower-triangular, unit-diagonal matrix such that its* NE *can be performed without changes of rows and such that the multipliers are positive. Then the* NE *of L^{-1} can also be carried out without changes of rows, and the multipliers are negatives of those of L, but, in general, used in a different order.*

Proof. The NE transforms L into I by

$$(5.3)\qquad \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & 0 & 1 & \\ & & & -l_{n,n-1} & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & & & & \\ -l_{21} & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & -l_{n1} & 1 \end{bmatrix} L = I,$$

which can be written

$$(5.4)\quad E_n(-l_{n,n-1})\,(E_{n-1}(-l_{n-1,n-2})E_n(-l_{n,n-2}))\cdots(E_2(-l_{21})\cdots E_n(-l_{n1}))\,L = I,$$

where the factors

$$(5.5)\qquad E_i(-l_{ik}) = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{ik} & 1 & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix}$$

are nontrivial because by hypothesis $l_{ik} > 0$. Here $-l_{ik}$ is the entry $(i, i-1)$ of the matrix.

Then

$$(5.6)\quad L^{-1} = E_n(-l_{n,n-1})\,(E_{n-1}(-l_{n-1,n-2})E_n(-l_{n,n-2}))\cdots(E_2(-l_{21})\cdots E_n(-l_{n1}))\,.$$

The elementary matrices $E_i(\alpha)$ defined by (5.5) with $\alpha \neq 0$ verify

$$(5.7)\qquad E_i^{-1}(\alpha) = E_i(-\alpha),$$

$$(5.8)\qquad E_i(\alpha)E_j(\beta) = E_j(\beta)E_i(\alpha) \quad \forall i,j \text{ such that } |i-j| \neq 1,$$

as is easily seen.

By (5.6) and (5.7)

$$(5.9)\qquad (E_n(l_{n1})\cdots E_2(l_{21}))\cdots(E_n(l_{n,n-2})E_{n-1}(l_{n-1,n-2}))\,E_n(l_{n,n-1})L^{-1} = I,$$

and by (5.8) this can be reordered in the form

$$(5.10)\quad E_n(l_{n1})\,(E_{n-1}(l_{n-1,1})E_n(l_{n2}))\cdots(E_2(l_{21})\cdots E_{n-1}(l_{n-1,n-2})E_n(l_{n,n-1}))\,L^{-1} = I,$$

that is,

$$(5.11)\qquad \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & 0 & 1 & \\ & & & l_{n1} & 1 \end{bmatrix} \begin{bmatrix} 1 & & & & \\ 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & l_{n-1,1} & 1 & \\ & & & l_{n2} & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & & & & \\ l_{21} & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & l_{n,n-1} & 1 \end{bmatrix} L^{-1} = I.$$

So we are in the same situation as in (5.2): The (i,k) multiplier of the NE of L^{-1} is the negative of the $(i, i-k)$ multiplier of L (compare (5.3) with (5.11)). □

Now we easily prove the following theorem, where U is the upper-triangular matrix obtained by the NE of A.

THEOREM 5.2. *A nonsingular matrix A is a strict γ-matrix if and only if all the pivots of the* NE *of A are positive and all the multipliers of the* NE *of U^T are negative.*

Proof. If A is a strict γ-matrix, then it is lowerly STP, and by [5, Prop. 4.4 and Lemma 2.6] the pivots of the NE of A are positive. By that elimination process we get the factorization $A = LDU$, where U is an upper-triangular, unit-diagonal matrix. Since A is a strict γ-matrix, the unicity of that factorization implies that U^{-1} is ΔSTP, and therefore $(U^T)^{-1} = (U^{-1})^T$ is ΔSTP also.

By [5, Cor. 5.5] we can apply Lemma 5.1 to the lower-triangular matrix $(U^T)^{-1}$, and we deduce that the multipliers of the NE of U^T are negative.

Conversely, if all the pivots of the NE of a nonsingular matrix A are positive, then from [5, Lemma 2.6] we get (4.1) and therefore A is lowerly STP.

On the other hand, if all the multipliers of the NE of U^T are negative, then by Lemma 5.1 all the multipliers of $(U^T)^{-1}$ are positive. As a consequence, all minors of $(U^T)^{-1}$ with consecutive initial columns and consecutive rows are positive. Therefore, by [2, Thm. 3.1] $(U^T)^{-1}$ is ΔSTP, and so, by Proposition 4.6, A is a strict γ-matrix. □

6. Remarks on strict γ-matrices in the diagonalization of oscillatory matrices. Oscillatory matrices are TP matrices A such that for some positive integer k A^k is STP. In [4, pp. 105–106], some properties of the eigenvalues and eigenvectors of those matrices are given: The eigenvalues λ_i are all simple, real, and positive, and the corresponding eigenvectors v_i can be chosen with certain sign variations in the sequence of their components. In [4, pp. 106–107], it was proved that A can be diagonalized

$$P^{-1}AP = \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\} = D \tag{6.1}$$

by a matrix P such that

$$\det P[\alpha|\beta] > 0 \quad \forall \alpha, \beta \in Q_{k,n} \text{ with } d(\beta) = 0,\ \beta_1 = 1. \tag{6.2}$$

That is, with our definitions and by Proposition 4.4, P is lowerly STP.

From (6.1) we get

$$A^T = (P^T)^{-1} D P^T, \tag{6.3}$$

and in [4, p. 107] it is also proved that P^T satisfies the property (6.2). Again by Proposition 4.4, $(P^{-1})^T$ is lowerly STP.

So by Proposition 4.6 P is a strict γ-matrix. In conclusion, an oscillatory matrix A can be decomposed in the form

$$A = P^{-1} D P, \tag{6.4}$$

where P is a strict γ-matrix and D is a diagonal matrix with positive diagonal entries.

From [1, Thm. 6.4] and its proof we can deduce the following result, which in some sense is a converse of (6.4).

Result. If A is a nonsingular real matrix with real eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0$ and if there exist strict γ-matrices P, N such that

$$A = P \cdot \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\} P^{-1}$$

and

$$A^T = N \cdot \operatorname{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\} N^{-1},$$

then there exists a positive integer k such that A^k is STP.

REFERENCES

[1] T. Ando, *Totally positive matrices*, Linear Algebra Appl., 90 (1987), pp. 165–219.

[2] C. W. Cryer, *LU-factorization of totally positive matrices*, Linear Algebra Appl., 7 (1973), pp. 83–92.

[3] ———, *Some properties of totally positive matrices*, Linear Algebra Appl., 15 (1976), pp. 1–25.

[4] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1959.

[5] M. Gasca and J. M. Peña, *Total positivity and Neville elimination*, Linear Algebra Appl., 44 (1992), pp. 25–44.

[6] S. Karlin, *Total positivity*, Stanford University Press, Stanford, CA, 1968.

[7] J. W. Rainey and G. J. Halbetler, *Tridiagonalization of completely nonnegative matrices*, Math. Comp., 26 (1972), pp. 121–128.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1141–1145, October 1993

018

ON THE PERTURBATION OF LU, CHOLESKY, AND QR FACTORIZATIONS*

G. W. STEWART†

Abstract. In this paper error bounds are derived for a first-order expansion of the LU factorization of a perturbation of the identity. The results are applied to obtain perturbation expansions of the LU, Cholesky, and QR factorizations.

Key words. perturbation theory, LU factorization, QR factorization, Cholesky factorization

AMS subject classification. 15A18

1. Introduction.

Let A be of order n, and suppose that the leading principal submatrices of A are nonsingular. Then A has an LU factorization

$$A = LU, \tag{1.1}$$

where L is lower triangular and U is upper triangular. The factorization is not unique; however, any other LU factorization must have the form

$$A = (LD)(D^{-1}U),$$

where D is a nonsingular diagonal matrix. Thus, if the diagonal elements of L (or U) are specified, the factorization is uniquely determined.

The purpose of this note is to establish a first-order perturbation expansion for the LU factorization of A along with bounds on the second-order terms. At least three authors have considered the perturbation of LU, Cholesky, and QR factorizations [1], [2], [4]. The chief difference between their papers and this one is that the former treat perturbations bounds for the decompositions in question, while here we treat the accuracy of a perturbation expansion.

Throughout this note $\|\cdot\|$ will denote a family of absolute, consistent matrix norms; i.e.,

$$|A| \le |B| \implies \|A\| \le \|B\|$$

and

$$\|AB\| \le \|A\|\|B\|$$

whenever the product AB is defined. Thus the bounds of this paper will hold for the Frobenius norm, the 1-norm, and the ∞-norm, but not for the 2-norm (for more on these norms see [3]).

2. Perturbation of the identity.

The heart of this note is the observation that the LU factorization of the matrix $I + F$, where F is small, has a simple perturbation expansion. Specifically, write

$$F = F_{\mathrm{L}} + F_{\mathrm{U}},$$

* Received by the editors February 28, 1992; accepted for publication (in revised form) March 9, 1992. This work was supported in part by Air Force Office of Sponsored Research contract AFOSR-87-0188.

† Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland 20742 (stewart@cs.umd.edu).

where F_{L} is strictly lower triangular and F_{U} is upper triangular. Then

$$(I + F_{\mathrm{L}})(I + F_{\mathrm{U}}) = I + F_{\mathrm{L}} + F_{\mathrm{U}} + F_{\mathrm{L}}F_{\mathrm{U}} = I + F + O(\|F\|^2), \tag{2.1}$$

and the product of the unit lower triangular matrix $I + F_{\mathrm{L}}$ and the upper triangular matrix $I + F_{\mathrm{U}}$ reproduces $I + F$ up to terms of order $\|F\|^2$. The following theorem shows that we can move these lower-order terms to the right-hand side of (2.1) to get an LU factorization of $I + F$.

THEOREM 2.1. *If*

$$\|F\| \le \frac{1}{4},$$

then there is a strictly lower triangular matrix G_{L} *and an upper triangular matrix* G_{U} *satisfying*

$$\|G_{\mathrm{L}} + G_{\mathrm{U}}\| \le \frac{\|F\|^2}{1 - 2\|F\| + \sqrt{1 - 4\|F\|}}$$

such that

$$(I + F_{\mathrm{L}} + G_{\mathrm{L}})(I + F_{\mathrm{U}} + G_{\mathrm{U}}) = I + F. \tag{2.2}$$

Proof. From (2.2) it follows that the perturbations G_{L} and G_{U} must satisfy

$$G_{\mathrm{L}} + G_{\mathrm{U}} = -(F_{\mathrm{L}}F_{\mathrm{U}} + F_{\mathrm{L}}G_{\mathrm{U}} + G_{\mathrm{L}}F_{\mathrm{U}} + G_{\mathrm{L}}G_{\mathrm{U}}).$$

Starting with $G_{\mathrm{L}}^0 = 0$ and $G_{\mathrm{U}}^0 = 0$, generate strictly lower triangular and upper triangular iterates according to the formula

$$G_{\mathrm{L}}^{k+1} + G_{\mathrm{U}}^{k+1} = -(F_{\mathrm{L}}F_{\mathrm{U}} + F_{\mathrm{L}}G_{\mathrm{L}}^k + G_{\mathrm{U}}^kF_{\mathrm{U}} + G_{\mathrm{L}}^kG_{\mathrm{U}}^k). \tag{2.3}$$

Because $\|\cdot\|$ is absolute,

$$\|G_{\mathrm{L}}^k\|, \|G_{\mathrm{U}}^k\| \le \|G_{\mathrm{L}}^k + G_{\mathrm{U}}^k\|.$$

Hence if we set $\phi = \|F\|$, $\gamma_0 = 0$, and define the sequence $\{\gamma_k\}$ by

$$\gamma_{k+1} = \phi^2 + 2\phi\gamma_k + \gamma_k^2, \qquad k = 0, 1, \ldots, \tag{2.4}$$

then $\|G_{\mathrm{L}}^k + G_{\mathrm{U}}^k\| \le \gamma_k$.

Now by graphing the right-hand side of (2.4), it is easy to see that if $\phi \le \frac{1}{4}$, then the sequence γ_k converges monotonically to

$$\gamma_* = \frac{\phi^2}{1 - 2\phi + \sqrt{1 - 4\phi}},$$

which is, therefore, an upper bound on $\|G_{\mathrm{L}}^k + G_{\mathrm{U}}^k\|$ for all k. It remains only to show that the sequence $G_{\mathrm{L}}^k + G_{\mathrm{U}}^k$ converges.

From (2.3) it follows that

$$\begin{aligned}(G_{\mathrm{L}}^{k+1} + G_{\mathrm{U}}^{k+1}) - (G_{\mathrm{L}}^k + G_{\mathrm{U}}^k) = {} & F_{\mathrm{L}}(G_{\mathrm{U}}^{k-1} - G_{\mathrm{U}}^k) + (G_{\mathrm{L}}^{k-1} - G_{\mathrm{L}}^k)F_{\mathrm{U}} \\ & + (G_{\mathrm{L}}^{k-1} - G_{\mathrm{L}}^k)G_{\mathrm{U}}^{k-1} + G_{\mathrm{L}}^k(G_{\mathrm{U}}^{k-1} - G_{\mathrm{U}}^k).\end{aligned}$$

Hence

$$\|(G_{\mathrm{L}}^{k+1} + G_{\mathrm{U}}^{k+1}) - (G_{\mathrm{L}}^{k} + G_{\mathrm{U}}^{k})\| \le 2(\phi + \gamma_*)\|(G_{\mathrm{L}}^{k} + G_{\mathrm{U}}^{k}) - (G_{\mathrm{L}}^{k-1} + G_{\mathrm{U}}^{k-1})\|.$$

If $2(\phi + \gamma_*) < 1$, which is certainly true if $\phi \le \frac{1}{4}$, then the series of differences is majorized by a geometric series, and the sequence converges.

There are some comments to be made on this theorem. In the first place the first-order expansion is particularly simple: Split F into its lower and upper triangular parts. We will take advantage of this simplicity in the next section, where we will derive perturbation expansions and asymptotic bounds for the LU, Cholesky, and QR factorizations.

The condition that $\|F\| \le \frac{1}{4}$ is perhaps too constraining, since the LU factorization of $I + F$ exists provided that $\|F\| < 1$. However, as $\|F\|$ approaches one, it is possible for the factors in the decomposition to grow arbitrarily, in which case the bounds on the second-order terms must also grow. Thus the more restrictive condition can be seen as the price we pay for bounds that do not explode.

As $\|F\|$ goes to zero, the bound quickly assumes the asymptotic form

$$\|G_{\mathrm{L}} + G_{\mathrm{U}}\| \lesssim \|F\|^2;$$

i.e., the order constant for the second-order terms is essentially one. If we write this in the form

$$\frac{\|G_{\mathrm{L}} + G_{\mathrm{U}}\|}{\|F\|} \lesssim \|F\|,$$

we see that the *relative* error in the first-order expansion is of the same order as the perturbation itself, with order constant one.

Finally, Theorem 2.1 treats an LU decomposition of $I+F$ in which L is unit lower triangular. In analyzing symmetric permutations, we may want to take $L = U^{\mathrm{T}}$. In this case, we may work with slightly different matrices, illustrated below for $n = 3$:

$$(2.5)\quad \hat{F}_{\mathrm{L}} = \begin{pmatrix} \frac{1}{2}f_{11} & 0 & 0 \\ f_{21} & \frac{1}{2}f_{22} & 0 \\ f_{31} & f_{32} & \frac{1}{2}f_{33} \end{pmatrix} \quad \text{and} \quad \hat{F}_{\mathrm{U}} = \begin{pmatrix} \frac{1}{2}f_{11} & f_{12} & f_{13} \\ 0 & \frac{1}{2}f_{22} & f_{23} \\ 0 & 0 & \frac{1}{2}f_{22} \end{pmatrix}.$$

If $\hat{G}_{\mathrm{L}}$ and $\hat{G}_{\mathrm{U}}$ are defined analogously, the proof of Theorem 2.1 goes through *mutatis mutandis*. For these matrices a useful inequality is

$$(2.6)\qquad \|\hat{F}_{\mathrm{L}}\|_{\mathrm{F}}, \|\hat{F}_{\mathrm{U}}\|_{\mathrm{F}} \le \frac{1}{\sqrt{2}}\|F\|_{\mathrm{F}},$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

3. Applications.

In this section we will apply the results of the previous section to get perturbation expansions for the LU, Cholesky, and QR decompositions. We will present only first-order terms, since bounds for the second-order terms can be derived from Theorem 2.1, and since the rate of convergence of these bounds to zero suggests that the first-order expansions will be satisfactory for all but the most delicate work. We will also derive asymptotic bounds for the first-order terms.

Our first application is to the problem we began with: the perturbation of the LU decomposition. Let A have the LU decomposition (1.1) and let $\tilde{A} = A + E$. Then

$$L^{-1}\tilde{A}U^{-1} = I + L^{-1}EU^{-1} \equiv I + F.$$

Let F_{L} and F_{U} be as in the last section. Then $I+F\cong(I+F_{\mathrm{L}})(I+F_{\mathrm{U}})$ is the first-order approximation to the LU factorization of $I+F$. It follows that

$$\tilde{A}\cong L(I+F_{\mathrm{L}})(I+F_{\mathrm{U}})U$$

is the first-order approximation to the LU factorization of $\tilde{A}$. Note that because F_{L} is unit lower triangular, this expansion preserves the scaling of the diagonal elements of L.

By taking norms we can derive the following asymptotic perturbation bound

$$\frac{\|\tilde{L}-L\|}{\|L\|}\lesssim\|L^{-1}\|\|U^{-1}\|\|A\|\frac{\|E\|}{\|A\|}\equiv\kappa_{\mathrm{LU}}(A)\frac{\|E\|}{\|A\|}. \tag{3.1}$$

Thus $\kappa_{\mathrm{LU}}(A)=\|L^{-1}\|\|U^{-1}\|\|A\|$ serves as a condition number for the LU decomposition of A. When A is square, this number is never less than the usual condition number $\kappa(A)=\|A\|\|A^{-1}\|$ and can be much larger. Bounds on the U factor can be derived similarly.

An unhappy aspect of the bound (3.1) is that it overestimates the perturbation of the leading part of the LU factorization. Specifically, if we partition

$$\begin{pmatrix}A_{11}&A_{12}\\A_{21}&A_{22}\end{pmatrix}=\begin{pmatrix}L_{11}&0\\L_{21}&L_{22}\end{pmatrix}\begin{pmatrix}U_{11}&U_{12}\\0&U_{22}\end{pmatrix},$$

then $A_{11}=L_{11}U_{11}$, and the condition number for this part of the factorization is $\kappa_{\mathrm{LU}}(A_{11})$, which is in general smaller than $\kappa_{\mathrm{LU}}(A)$. The perturbation in L_{21} can then be estimated from the equation $L_{21}=A_{21}U_{11}^{-1}$.[1]

If A is symmetric and positive definite, then A has the Cholesky factorization

$$A=R^{\mathrm{T}}R,$$

where R is upper triangular. Let $\tilde{A}=A+E$, where E is symmetric. Setting $F=R^{-\mathrm{T}}ER^{-1}$ and defining $\hat{F}_{\mathrm{U}}$ as in (2.5), we have

$$\tilde{R}\cong(I+\hat{F}_{\mathrm{U}})R.$$

By (2.6) and the consistency of the 2-norm with the Frobenius norm, we have

$$\frac{\|\tilde{R}-R\|_{\mathrm{F}}}{\|R\|_2}\lesssim\frac{1}{\sqrt{2}}\|R^{-\mathrm{T}}\|_2\|R^{-1}\|_2\|E\|_{\mathrm{F}}=\frac{\kappa_2(A)}{\sqrt{2}}\frac{\|E\|_{\mathrm{F}}}{\|A\|_2},$$

where $\kappa_2(A)=\|A\|_2\|A^{-1}\|_2$ is the usual condition number in the 2-norm.

Finally, let A, now rectangular, be of full-column rank, and consider the QR factorization

$$A=QR,$$

[1] An alternative approach is to set

$$\check{L}=\begin{pmatrix}L_{11}&0\\L_{21}&I\end{pmatrix},$$

so that

$$\check{L}^{-1}\begin{pmatrix}A_{11}\\A_{21}\end{pmatrix}U_{11}^{-1}=\begin{pmatrix}I\\0\end{pmatrix}\equiv J.$$

The proof of Theorem 2.1 can easily be adapted to give a bound on the perturbation of the LU factorization of a perturbation of J and hence on the perturbation of the LU factorization of $\left(\begin{smallmatrix}A_{11}\\A_{21}\end{smallmatrix}\right)$.

where Q has orthonormal columns and R is upper triangular with positive diagonal elements. The key to the derivation of the bounds is the equation

$$A^{\mathrm{T}}A = R^{\mathrm{T}}R;$$

i.e., R is the Cholesky factor of $A^{\mathrm{T}}A$.

As usual, let $\tilde{A} = A + E$, and let E_A be the orthogonal projection of E onto the column space of A. Then $A^{\mathrm{T}}E = A^{\mathrm{T}}E_A$. It follows that

$$\tilde{A}^{\mathrm{T}}\tilde{A} \cong A^{\mathrm{T}}A + A^{\mathrm{T}}E_A + E_A^{\mathrm{T}}A \equiv A^{\mathrm{T}}A + F.$$

Hence, with $\hat{F}_{\mathrm{U}}$ as above, we have

$$\tilde{R} \cong (I + \hat{F}_{\mathrm{U}})R.$$

In particular,

$$\frac{\|\tilde{R} - R\|_{\mathrm{F}}}{\|R\|_2} \lesssim \sqrt{2}\kappa_2(A)\frac{\|E_A\|_{\mathrm{F}}}{\|A\|_2},$$

where $\kappa_2(A) = \|R\|_2\|R^{-1}\|_2$. Since $\tilde{Q} = \tilde{A}\tilde{R}^{-1}$, we have

$$\tilde{Q} \cong Q(I - \hat{F}_{\mathrm{U}}) + ER^{-1},$$

from which it follows that

$$\|\tilde{Q} - Q\|_{\mathrm{F}} \lesssim \kappa_2(A)\frac{\sqrt{2}\|E_A\|_{\mathrm{F}} + \|E\|_{\mathrm{F}}}{\|A\|_2}. \tag{3.2}$$

Asymptotically, the bounds derived in this section agree with the bounds in [1], [4], with the exception of (3.2), which is a little sharper owing to the presence of $\|E_A\|_{\mathrm{F}}$.

Acknowledgment. This paper has been much improved by the suggestions of Jim Demmel and Nick Higham.

REFERENCES

[1] A. Barrland, *Perturbation bounds for the LDL^{H} and the* LU *factorizations*, BIT, 31 (1991), pp. 358–363.

[2] G. W. Stewart, *Perturbation bounds for the* QR *factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.

[3] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[4] J.-G. Sun, *Perturbation bounds for the Cholesky and* QR *factorizations*, BIT, 31 (1991), pp. 341–352.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1146–1151, October 1993

019

A NOTE ON GENERALIZED VANDERMONDE DETERMINANTS*

RANDOLPH P. FLOWE† AND GARY A. HARRIS†

Abstract. This note contains elementary derivations of explicit formulae for the determinants of two types of matrices which are generalized versions of that which is commonly known as the Vandermonde matrix.

Key words. generalized Vandermonde matrices, Hermite interpolation, linear recursion relation, degree of a Grassmann manifold

AMS subject classifications. primary 15A15; secondary 41A05, 65F40, 65Q05

1. Introduction. In this note we present an elementary derivation of explicit formulae for the determinants of two types of matrices that occur quite frequently in the literature. We view both types as generalized Vandermonde matrices. A matrix of the first type, which we will always denote by M, is constructed as follows. Let $z_1, \ldots, z_\ell$ be distinct complex numbers with associated positive integers $m_1, \ldots, m_\ell$. Let $k \doteq m_1 + \cdots + m_\ell$ and, for each $1 \leq j \leq \ell$, let c_j denote the column vector $(1, z_j, z_j^2, \ldots, z_j^{k-1})^T$. Then M denotes the k-by-k matrix

$$M \doteq [c_1, c_1', \ldots, c_1^{(m_1-1)}; \ldots; c_\ell, c_\ell', \ldots, c_\ell^{(m_\ell-1)}]. \tag{1.1}$$

A superscript on c_j in (1.1) denotes the number of applications of $\frac{\partial}{\partial z_j}$ to c_j.

THEOREM 1.1.

$$\det\ M = \left(\prod_{j=1}^{\ell} [0!1! \ldots (m_j - 1)!] \right) \prod_{i>j} (z_i - z_j)^{m_i m_j}.$$

Such matrices occur naturally in the process of Hermite interpolation [4]. Notice in case $\ell = k$ (so each $m_j = 1$) that M is the Vandermonde matrix, and Theorem 1.1 yields the well-known formula $\det\ M = \prod_{i>j}(z_i - z_j)$.

Matrices of the second type that we wish to consider are constructed in the same fashion as M above, except a superscript on c_j denotes the number of applications of the operator

$$D_j \doteq z_j \frac{\partial}{\partial z_j} \tag{1.2}$$

to c_j. To avoid confusion, henceforth, we let $\tilde{M}$ denote such a matrix. Thus

$$\tilde{M} \doteq [c_1, D_1 c_1, \ldots, D_1^{m_1-1} c_1; \ldots; c_\ell, D_\ell c_\ell, \ldots, D_\ell^{m_\ell-1} c_\ell]. \tag{1.3}$$

THEOREM 1.2.

$$\det\ \tilde{M} = \left(\prod_{j=1}^{\ell} z_j^{m_j(m_j-1)/2} \right) \det\ M.$$

* Received by the editors June 10, 1991; accepted for publication (in revised form) February 16, 1992.

† Department of Mathematics, Texas Tech University, Mail Stop 1042, Lubbock, Texas 79409 (magah@ttacs1.ttu.edu).

Matrices of the form $\tilde{M}$ are key to finding general solutions to linear recursion relations [1]. Notice that $\tilde{M}$ in (1.3) is again the Vandermonde matrix if $\ell = k$. Also notice that Theorem 1.1 implies det $M \neq 0$ (since $z_i \neq z_j$ for $i \neq j$), and Theorem 1.2 implies det $\tilde{M} \neq 0$ if each $z_j \neq 0$. (det $M \neq 0$ is needed to show that Hermite interpolation actually works [4], [3], and det $\tilde{M} \neq 0$ is needed to prove the existence of a general solution to a given linear recursion relation [1].)

As indicated previously, the purpose of this note is to provide a particularly elementary and, we hope, interesting derivation of the formulae in Theorems 1.1 and 1.2. Theorem 1.2 follows from an elementary reduction argument given in Lemma 3.1. The formula in Theorem 1.1 is obtained by an inductive procedure completely analogous to that which is often used to obtain the earlier-mentioned formula for the Vandermonde determinant. What makes our derivation interesting is our use of the Leibniz rule for differentiating determinants to construct a tree (see Fig. 2.2) which has encoded into it all the pertinent information about the derivatives of det M. An important feature of this process is our ability to count the total number of distinct branches in such a tree (see Lemma 2.1). An intriguing feature of this tree is that the number of branches in it is equal to the degree of a certain Grassmann manifold (see Remark 4.1). This seems unlikely to be only coincidence, and we are currently studying facets of the Schubert calculus in the hope of eventually understanding the relation between these a priori unrelated concepts.

2. Proof of Theorem 1.1. Define the polynomial p by

$$p(z_1) \doteq \det\ M,$$

thinking of z_1 as variable and $z_2, \ldots, z_\ell$ as constants. Then

$$\begin{aligned} p'(z_1) &= \det\ [c_1, c_1', \ldots, c_1^{(m_1-2)}, c_1^{(m_1)}; *], \\ p''(z_1) &= \det\ [c_1, c_1', \ldots, c_1^{(m_1-3)}, c_1^{(m_1-1)}, c_1^{(m_1)}; *] \\ &\quad + \det\ [c_1, c_1', \ldots, c_1^{(m_1-2)}, c_1^{(m_1+1)}; *], \end{aligned}$$

and so forth. Here $*$ is used to denote the columns which are constant with respect to z_1. We need to pursue this differentiation process to its conclusion, so we shorten the notation by using only the differentiation superscripts on the c_1 columns. Thus $|0, 1, \ldots, m_1 - 1|$ corresponds to $p(z_1)$, $|0, 1, \ldots, m_1 - 2, m_1|$ to $p'(z_1)$, $|0, 1, \ldots, m_1 - 3, m_1 - 1, m_1| + |0, 1, \ldots, m_1 - 2, m_1 + 1|$ to $p''(z_1)$, etc. In general, for integers $0 \leq i_1 < i_2 < \cdots < i_{m_1}$, $|i_1, i_2, \ldots, i_{m_1}|$ corresponds to det $[c_1^{(i_1)}, c_1^{(i_2)}, \ldots, c_1^{(i_{m_1})}; *]$. Observe that

$$\frac{\partial}{\partial z_1} |i_1, \ldots, i_{m_1}| = \sum_{j=1}^{m_1} |i_1, \ldots, i_{j-1}, i_j + 1, i_{j+1}, \ldots, i_{m_1}|. \tag{2.1}$$

We record some observations for future use.

(i) If $i_{m_1} \geq k$, then $|i_1, i_2, \ldots, i_{m_1}|$ vanishes.

(ii) If for any $1 \leq j \leq m_1 - 1$, $i_{j+1} = i_j + 1$, then $|i_1, \ldots, i_{j-1}, i_j + 1, i_{j+1}, \ldots, i_{m_1}|$ vanishes.

(iii) It requires m_1 differentiations of $\rho(z_1)$ to obtain $|i_1 + 1, i_2 + 1, \ldots, i_{m_1} + 1|$.

We now construct a diagram which describes completely the process of differentiating $p(z_1)$. Place $p(z_1)$ at the top of the diagram. Each row of the diagram is obtained by differentiating that which is immediately above it and ordering the determinants left to right as j goes from 1 to m_1 in (2.1). Any duplicate determinants are collected into a single term (with integer coefficient) in the position of that which first occurs on the left. Figure 2.1 is the diagram so constructed in case $k = 6$ and $m_1 = 3$.

$$\begin{gathered}
|0,1,2| \\
|0,1,3| \\
|0,2,3| + |0,1,4| \\
|1,2,3| + 2|0,2,4| + |0,1,5| \\
3|1,2,4| + 2|0,3,4| + 3|0,2,5| \\
5|1,3,4| + 6|1,2,5| + 5|0,3,5| \\
5|2,3,4| + 16|1,3,5| + 5|0,4,5| \\
21|2,3,5| + 21|1,4,5| \\
42|2,4,5| \\
42|3,4,5|
\end{gathered}$$

FIG. 2.1. *det M and its derivatives in case $k = 6, m_1 = 3$.*

Such a diagram gives rise to a tree. Simply connect a determinant in any row to those derived from it in the next row. Now the integer coefficient counts the number of distinct paths (branches) from the top of the tree to the given determinant. Figure 2.2 shows the tree obtained from the diagram in Fig. 2.1.

Note that the tree in Fig. 2.2 has a total of 42 distinct branches. In general we let $N(k, m_1)$ denote the total number of distinct branches in such a tree.

The pertinent properties of any such tree (obtained from observations (i), (ii), and (iii)) are the following.

Property 1. Letting the top row be row 0, row j corresponds to $p^{(j)}(z_1)$.

Property 2. In each row the leading entry of the leftmost term is greater than or equal to the leading entries of the other terms in that row.

Property 3. For any nonnegative integer ν, the leftmost term in row νm_1 is $|\nu, \nu+1, \ldots, \nu+m_1-1|$, and $\nu + m_1 - 1$ is strictly less than the last entry of any other term in this row. Also, all terms in all rows above row νm_1 have leading entry strictly less than ν.

Property 4. If $2 \leq j \leq \ell$, then each term in each row above row $m_1 m_j$ contains at least one entry equal to $0, 1, 2, \ldots,$ or $m_j - 1$.

Property 5. The tree ends with row $(k - m_1)m_1$. And that row is ${}_{N(k,m_1)}|k - m_1, \ldots, k-1|$.

Recall that $N(k, m_1)$ is the number of distinct branches in the tree. It is also the number of determinants $|k - m_1, \ldots, k - 1|$ generated in the diagram. It is a number we explicitly compute in Lemma 2.1 (proof in §4).

It follows from Properties 1 and 5 that $\deg p \leq (k - m_1)m_1$. Property 4 implies that $z_j, 2 \leq j \leq \ell$, is a root of p with multiplicity at least $m_1 m_j$. Thus

$$p(z_1) = c_1 \prod_{j=2}^{\ell} (z_1 - z_j)^{m_1 m_j} = (-1)^{(k-m_1)m_1} c_1 \prod_{j=2}^{\ell} (z_j - z_1)^{m_1 m_j}. \tag{2.2}$$

Since $\sum_{j=2}^{\ell} m_1 m_j = (k - m_1)m_1 \geq \deg\ p$, it follows that c_1 in (2.2) is independent

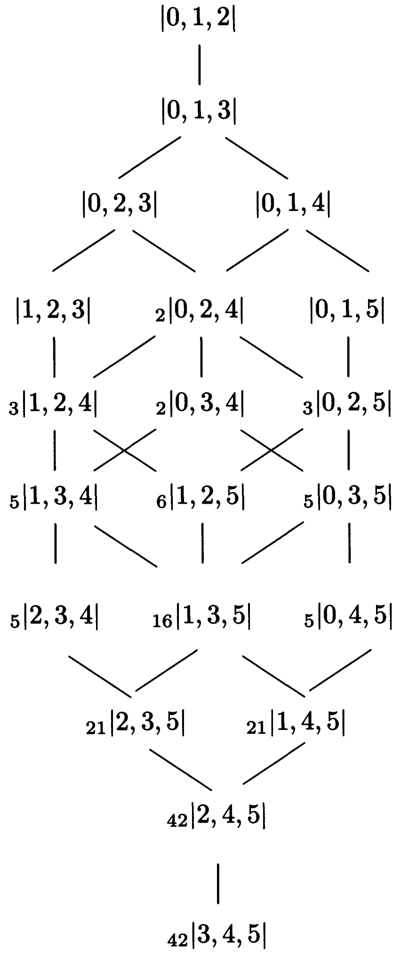


FIG. 2.2. *Tree associated with det M and its derivatives.*

of z_1 and

$$c_1 = \frac{1}{[(k-m_1)m_1]!} p^{((k-m_1)m_1)}(0). \tag{2.3}$$

To compute c_1 we introduce a bit more notation. For $1 \le j \le \ell$ let $B_j \doteq [c_j, c_j', \dots, c_j^{(m_j-1)}]$. So $p(z_1) = \det\,[B_1(z_1); B_2; \dots; B_\ell]$ and

$$p^{((k-m_1)m_1)}(z_1) = N(k, m_1)|k-m_1, k-m_1+1, \dots, k-1| =$$

$$N(k,m_1)\det\left[\begin{array}{ccccc} 0 & 0 & \dots\ \dots & 0 & \\ \vdots & \vdots & & \vdots & \\ 0 & & \dots\ \dots & \vdots & \\ (k-m_1)! & 0 & \dots\ \dots & \dots; & B_2;\dots;B_\ell \\ & (k-m_1+1)! & \dots\ \dots & & \\ & & \ddots & & \\ & \star & \ddots & 0 & \\ & & & (k-1)! & \end{array}\right]$$

where $\star$ consists of z_1's to positive powers. Expanding the determinant by the columns of the z_1 block we obtain

$$p^{((k-m_1)m_1)}(0) = N(k,m_1)(-1)^{(k+m_1)m_1}(k-m_1)!\dots(k-1)!\det\,[\hat{B}_2;\dots;\hat{B}_\ell], \tag{2.4}$$

where each $\hat{B}_j$ is obtained from B_j by deleting the last m_1 rows. Substituting (2.4) and (2.3) into (2.2) yields

$$p(z_1) = N(k,m_1)\left\{\frac{(k-m_1)!\dots(k-1)!}{[(k-m_1)m_1]!}\right\}\det\,[\hat{B}_2;\dots;\hat{B}_\ell]\prod_{j=2}^{\ell}(z_j-z_1)^{m_1m_j}. \tag{2.5}$$

The last calculation we need follows in Lemma 2.1.

LEMMA 2.1. (*Proof in* §4.)

$$N(k,m_1) = \frac{[0!1!\dots(m_1-1)!][(k-m_1)m_1]!}{(k-m_1)!\dots(k-1)!}. \tag{2.6}$$

It follows from (2.5) and (2.6) that

$$\det\, M = [0!1!\dots(m_1-1)!]\det\,[\hat{B}_2;\dots;\hat{B}_\ell]\prod_{j=2}^{\ell}(z_j-z_1)^{m_1m_j}.$$

Note that $[\hat{B}_2;\dots;\hat{B}_\ell]$ is a $k-m_1$ by $k-m_1$ matrix of the same form as M. Theorem 1.1 follows by induction. □

3. Proof of Theorem 1.2. Theorem 1.2 follows by applying the proof of the following lemma to each block $\tilde{B}_j \doteq [c_j, D_jc_j;\dots,D_j^{m_j-1}c_j]$ of the matrix $\tilde{M} \doteq [\tilde{B}_1;\dots;\tilde{B}_\ell]$.

LEMMA 3.1. *Define the operator D by*

$$D \doteq z\frac{\partial}{\partial z}.$$

For m a positive integer and $c(z)$ an arbitrary column vector of length m,

$$\det\,[c(z), Dc(z), \ldots, D^{m-1}c(z)] = z^{\frac{(m-1)m}{2}} \det\,[c(z), c'(z), \ldots, c^{(m-1)}(z)]. \tag{3.1}$$

Proof. Using $D(z^n c^{(n)}) = z^{n+1}c^{(n+1)} + nz^n c^{(n)}$ we get $\det\,[c, Dc, \ldots, D^{m-1}c] = \det\,[c, zc', z^2c''+zc', z^3c'''+3z^2c''+zc', \ldots, z^{m-1}c^{(m-1)}+\cdots+zc']$. Adding appropriate multiples of column 2 to succeeding columns, then appropriate multiples of column 3 to succeeding columns, etc., and finally factoring the common powers of z from each column, we obtain (3.1). □

4. Proof of Lemma 2.1.

The key to proving Lemma 2.1 is the observation that $N(k, m_1)$ is a property of the tree, dependent only on $|0, \ldots, m_1 - 1|$ and k, and independent of $B_2, \ldots, B_\ell$. Thus we may assume $B_2, \ldots, B_\ell$ to be especially nice. In particular assume $m_j = 1$ for each $2 \le j \le \ell$. Thus for $2 \le j \le \ell$ the z_j's are distinct and $B_j = (1, z_j, \ldots, z_j^{k-1})^T$. Let

$$A \doteq [B_1(z_1); B_2; \ldots; B_\ell]$$

and

$$B \doteq [B_2; \ldots; B_\ell; B_1(z_1)].$$

We know $\det\, A = \pm \det\, B$. Inducting on (2.5) and observing that $N(\hat{k}, 1) = 1$ for all $\hat{k}$ we get

$$\det\, A = N(k, m_1) \left\{ \frac{(k-m_1)! \ldots (k-1)!}{[(k-m_1)m_1]!} \right\} \prod_{i>j} (z_i - z_j)^{m_i m_j}. \tag{4.1}$$

Similar use of (2.5) and the observation that $N(m_1, m_1) = 1$ yields

$$\det\, B = \pm[0!1! \ldots (m_1 - 1)!] \prod_{i>j} (z_i - z_j)^{m_i m_j}. \tag{4.2}$$

(The "±" occurs in (4.2) because we are forcing the ordering on the z_j's to be the same as that used in (4.1).) Setting (4.1) = ± (4.2) and solving for $N(k, m_1)$ yields (2.6). □

Remark 4.1. The right side of (2.6) is a familiar expression in enumerative geometry. It is the degree of the complex Grassmann manifold of $(m_1 - 1)$ planes in projective $(k-1)$ space [2]. Why the number of branches in a tree constructed as above should equal this degree is a mystery that merits further study. However, such study promises to exceed the elementary tenants pervasive in this note.

Acknowledgments. We are especially grateful to our two colleagues, Clyde F. Martin and Alex Wang, for their helpful comments and suggestions regarding the contents of this note.

REFERENCES

[1] R. P. Grimaldi, *Discrete and Combinatorial Mathematics*, 2nd Ed., Addison–Wesley, Reading, MA, 1989.

[2] S. L. Kleiman, *Problem* 15. *Rigorous foundations of Schubert's enumerative calculus*, Proc. Symposia in Pure Mathematics, 28 (1976), pp. 445–489.

[3] C. Martin, J. Miller, and K. Pearce, *Sums of exponentials with polynomial coefficients*, Appl. Math. Comput., 40 (1990), pp. 33–39.

[4] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, R. Bartels, W. Gautschi, and C. Witzgall, translators, Springer-Verlag, New York, 1980.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1152–1167, October 1993

020

THE HADAMARD OPERATOR NORM OF A CIRCULANT AND APPLICATIONS*

ROY MATHIAS†

Abstract. Let M_n be the space of $n \times n$ complex matrices and let $\|\cdot\|_\infty$ denote the spectral norm. Given matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ of the same size, define their *Hadamard product* to be $A \circ B = [a_{ij}b_{ij}]$. Define the Hadamard operator norm of $A \in M_n$ to be

$$|||A|||_\infty = \max\{\|A \circ B\|_\infty : \|B\|_\infty \leq 1\}.$$

It is shown that

$$|||A|||_\infty = \mathrm{tr}|A|/n \tag{1}$$

if and only if

$$|A| \circ I = |A^*| \circ I = (\mathrm{tr}|A|/n)I. \tag{2}$$

It is shown that (2) holds for generalized circulants and hence that the Hadamard operator norm of a generalized circulant can be computed easily. This allows us to compute or bound

$$|||[\mathrm{sign}(j-i)]_{i,j=1}^n|||_\infty, \quad |||[(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)]_{i,j=1}^n|||_\infty, \quad |||T_n|||_\infty,$$

where T_n is the $n \times n$ matrix with ones on and above the diagonal and zeros below, and related quantities. In each case the norms grow like $\log n$.

Using these results upper and lower bounds are obtained on quantities of the form

$$\sup\{\|\,|A| - |B|\,\|_\infty : \|A - B\|_\infty \leq 1, A, B \in M_n\}$$

and

$$\sup\{\|\,|A|B - B|A|\,\|_\infty : \|AB - BA\|_\infty \leq 1,\ A, B \in M_n,\ A = A^*\}.$$

The authors also indicate the extent to which the results generalize to all unitarily invariant norms, characterize the case of equality in a matrix Cauchy–Schwarz Inequality, and give a counterexample to a conjecture involving Hadamard products.

Key words. Hadamard product, circulant, triangular truncation, commutator, Cauchy–Schwarz inequality, matrix absolute value

AMS subject classifications. 15A60, 15A45, 15A57, 47A55, 47B47

1. Introduction.

Let M_n denote the space of $n \times n$ complex matrices, and let R_+^n denote the cone of n-vectors with positive components. Given $A \in M_n$ we define $|A| \equiv (A^*A)^{1/2}$, the *Schatten p-norm* of A to be

$$\|A\|_p \equiv (\mathrm{tr}|A|^p)^{1/p} \quad \text{for } 1 \leq p < \infty$$

and $\|A\|_\infty$ to be the spectral norm of A (i.e., $\|A\|_\infty^2$ is the spectral radius of A^*A). Given matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ of the same size we define their *Hadamard product* to be $A \circ B = [a_{ij}b_{ij}]$. Given a norm $\|\cdot\|$ on M_n we define the corresponding Hadamard operator norm $|||\cdot|||$ on M_n by

$$|||A||| \equiv \max\{\|A \circ B\| : \|B\| \leq 1\}. \tag{1.1}$$

We are most interested in $|||\cdot|||_\infty$ and $|||\cdot|||_1$. Given Hermitian $A, B \in M_n$ we use $A \geq B$ to mean that $A - B$ is positive semidefinite.

* Received by the editors July 29, 1991; accepted (in revised form) January 8, 1992.

† Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23187 (na.mathias@na-net.ornl.gov). This research was supported by a Summer Research grant from the College of William and Mary.

There are few classes of matrices for which $|||\cdot|||_\infty$ is easy to compute. It is known that

$$(1.2)\qquad \begin{array}{ll} |||A|||_\infty = \max\{a_{ii} : i = 1, \dots, n\} & \text{if } A \text{ is positive semidefinite,} \\ |||A|||_\infty = 1 & \text{if } A \text{ is unitary.} \end{array}$$

In §2 we define a class of matrices that generalizes the circulants, and show that if A is such a generalized circulant then

$$|||A|||_\infty = \|A\|_1/n = \frac{1}{n}\sum_{i=1}^{n} |\lambda_i(A)|$$

and that $\lambda_i(A)$, the eigenvalues of A, are very easy to compute. We also determine the case of equality in a matrix Cauchy–Schwarz Inequality.

In §3 we will apply the results in §2 to determine the Hadamard operator norm of the $n \times n$ matrices

$$(1.3)\qquad S_n = [\operatorname{sign}(j-i)],$$

$$(1.4)\qquad \hat{S}_n = S_n + I,$$

and to bound the Hadamard operator norm of

$$(1.5)\qquad T_n = \text{ the } n \times n \text{ strictly upper triangular matrix of ones } (t_{ij} = 0 \text{ if } i \geq j \text{ and } t_{ij} = 1 \text{ if } i < j),$$

$$(1.6)\qquad \hat{T}_n = T_n + I.$$

We will also consider matrices of the form $[(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)]$ where λ_i are positive numbers. Let $J_n \in M_n$ be the matrix of ones.

In §4 we will use the bounds in §3 to prove some inequalities involving commutators, the matrix absolute value, and the eigenvalues of arbitrary perturbations of Hermitian matrices.

In §5 we will use the results of §2 to give a counterexample to the natural conjecture that if A is real and there is a real unitary U such that $A \circ U$ is entrywise nonnegative then

$$\max\{\|A \circ B\|_\infty : \|B\|_\infty \leq 1\}$$

is attained at a matrix B such that $A \circ B$ is entrywise nonnegative.

2. General results. In this section we characterize the matrices $A \in M_n$ for which $|||A|||_\infty = \|A\|_1/n$ and show that the circulants are in this class.

Our first result is perhaps well known but appears not to have been formally stated before.

LEMMA 2.1. *Let $A \in M_n$. Then*

$$|||A|||_\infty = |||A|||_1.$$

Proof. For any $X, Y, Z \in M_n$ we have

$$(2.1)\qquad \operatorname{tr}(X \circ Y)Z = \operatorname{tr}(X \circ Z^T)Y^T.$$

It is a simple exercise to verify this directly. Now by the duality (with respect to the inner product trAB^*) between the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ we have

$$\begin{aligned}|||A|||_\infty &= \max\{\|A\circ B\|_\infty : \|B\|_\infty \le 1\}\\ &= \max\{|\mathrm{tr}(A\circ B)C^*| : \|B\|_\infty \le 1, \|C\|_1 \le 1\}\\ &= \max\{|\mathrm{tr}(A\circ C^{*T})B^T| : \|B\|_\infty \le 1, \|C\|_1 \le 1\}\\ &= \max\{\|A\circ C^{*T}\|_1 : \|C\|_1 \le 1\}\\ &= \max\{\|A\circ C\|_1 : \|C\|_1 \le 1\}\\ &= |||A|||_1.\end{aligned}$$

We have also used $\|B^T\|_\infty = \|B\|_\infty$ and $\|C^{*T}\|_1 = \|C\|_1$. □

The next result is a very useful characterization of matrices with $|||A|||_\infty \le 1$ due to Haagerup (unpublished). See [2] or [13, Thm. 3.1] for a proof.

LEMMA 2.2. *Let $A \in M_n$. Then $|||A|||_\infty \le 1$ if and only if there are matrices $P, Q \in M_n$ with main diagonal entries at most* 1 *such that*

$$\begin{pmatrix} P & A \\ A^* & Q \end{pmatrix}$$

is positive semidefinite.

The following technical lemma will only be used to prove the "only if" part of Lemma 2.4 which is not used in the rest of the paper.

LEMMA 2.3. *Let $\|\cdot\|$ be a unitarily invariant norm on M_n such that*

$$A \ge B \ge 0 \quad \text{and} \quad A \ne B \quad \textit{imply} \quad \|A\| > \|B\|. \tag{2.2}$$

Let $A, B, C \in M_n$ be such that

$$\begin{pmatrix} A & C \\ C^* & B \end{pmatrix} \ge 0. \tag{2.3}$$

Then

$$\|C\|^2 \le \|A\|\ \|B\| \tag{2.4}$$

with equality if and only if $A = \alpha|C^*|$ and $B = \alpha^{-1}|C|$ for some $\alpha > 0$.

Note that (2.3) implies (2.4) for any unitarily invariant norm regardless of whether (2.2) holds or not.

Proof. Let $A, B, C \in M_n$ satisfy (2.3). Then it is well known that (2.4) holds (see e.g., [11, Thm. 2.3]). It is clear that if $A = |C^*|$ and $B = |C|$ then we have equality in (2.4). So it is sufficient to prove the "only if" part.

Now suppose that equality holds in (2.4). Without loss of generality we may assume that $\|A\| = \|B\|$. Let $U \in M_n$ be a unitary matrix such that $UC = |C|$. Then

$$\begin{pmatrix} UAU^* & |C| \\ |C| & B \end{pmatrix} \ge 0. \tag{2.5}$$

For positive semidefinite matrices X, Y let $X\#Y$ denote their geometric mean (see [1] for definition). Let $S = (UAU^* + B)/2$ and $D = (UAU^* - B)/2$ and let $G = (UAU^*)\#B$. Then by (2.5) together with the alternate characterization of the geometric mean on [1, p. 33] for the second inequality and [1, Eq. (11)] for the second equality we have

$$0 \le |C| \le G = (UAU^*)\#B = S\#(S - DS^\dagger D) \tag{2.6}$$

($S^\dagger$ is the Moore–Penrose generalized inverse of S). Again by the alternate definition of the geometric mean we have

$$\begin{pmatrix} S & G \\ G & S - DS^\dagger D \end{pmatrix} \ge 0. \tag{2.7}$$

So, using (2.6) for the first inequality and (2.7) for the second we have

$$\begin{aligned} \|C\|^2 = \| \, |C| \, \|^2 \ \le \|G\|^2 &\le \|S\| \, \|S - DS^\dagger D\| \\ &\le \|S\|^2 \\ &= \|(UAU^* + B)/2\|^2 \\ &\le [(\|A\| + \|B\|)/2]^2 \\ &= \|A\| \, \|B\| \\ &= \|C\|^2. \end{aligned}$$

Thus we must have equality throughout, and in particular, in the third inequality. That is $\|S\| = \|S - DS^\dagger D\|$, which by (2.2) implies that $DS^\dagger D = 0$ which in turn implies $D = 0$ (since the range of D is contained in that of S). So we have shown that $B = UAU^*$. Thus (2.5) is

$$\begin{pmatrix} B & |C| \\ |C| & B \end{pmatrix} \ge 0,$$

and hence $B \ge |C|$. So if $B \ne |C|$ then by (2.2) we would have $\|B\| > \| \ C \ \| = \|C\|$ which is not possible since we have equality in (2.4). Thus $B = |C|$, and now it is easy to show that $A = |C^*|$. □

LEMMA 2.4. *Let $A \in M_n$. Then*

$$|||A|||_\infty \ge \frac{1}{n}\mathrm{tr}|A| = \frac{1}{n}\|A\|_1. \tag{2.8}$$

Equality holds in (2.8) *if and only if*

$$|A| \circ I = |A^*| \circ I = \left(\frac{1}{n}\mathrm{tr}|A|\right) I \tag{2.9}$$

or, equivalently, if and only if the main diagonal entries of $|A|$ and $|A^*|$ are equal to $\|A\|_1/n$. If (2.9) *holds and U is a contraction such that $UA = |A|$, then $|||A|||_\infty = \|A \circ U^T\|_\infty$.*

Proof. Take any $A \in M_n$. Let U be a contraction such that $UA = |A|$. Then

$$|||A|||_\infty \ge \|A \circ U^T\|_\infty \ge (e_n^*/\sqrt{n})(A \circ U^T)(e_n/\sqrt{n}) = \mathrm{tr}(A \circ U^T)J/n = \frac{1}{n}\mathrm{tr}UA = \frac{1}{n}\mathrm{tr}|A|,$$

which gives (2.8). (e_n is the vector of 1's; we have used $e_n e_n^* = J_n$ and (2.1) for the second equality.)

The matrix

$$\begin{pmatrix} |A^*| & A \\ A^* & |A| \end{pmatrix} \tag{2.10}$$

is positive semidefinite and if (2.9) holds then the main diagonal entries are no greater than $\|A\|_1/n$ (in fact they are all equal to this) so by Lemma 2.2 $|\!|\!|A|\!|\!|_\infty \leq \|A\|_1/n$. (So we must have equality in (2.8).)

Now let us prove the "only if" part. Assume that $|\!|\!|A|\!|\!|_\infty = \|A\|_1/n$. Then there are matrices $P, Q \in M_n$ such that

$$\begin{pmatrix} P & A \\ A^* & Q \end{pmatrix} \geq 0 \tag{2.11}$$

and $\max p_{ii} = \max q_{ii} \leq \mathrm{tr}|A|/n$, and so $\|P\|_1 = \mathrm{tr}P \leq \|A\|_1$ and $\|Q\|_1 = \mathrm{tr}Q \leq \|A\|_1$. In view of (2.4) we must have $p_{ii} = q_{ii} = \mathrm{tr}|A|/n, \quad i = 1, \ldots, n$ and hence $\|P\|_1 = \|Q\|_1 = \|A\|_1$. So by the "only if" part of Lemma 2.3 $P = |A^*|$ and $Q = |A|$. Thus (2.9) follows from the fact that both P and Q have constant main diagonal entries. □

COROLLARY 2.5. *Let $A \in M_n$. Then*

$$|\!|\!|A|\!|\!|_\infty \geq \max\{\|B\|_1/k : B \in M_k \textit{ is a sub-matrix of } A, k = 1, \ldots, n\}.$$

The unitaries are an obvious class of matrices that satisfy (2.9). We now define the class of generalized circulants and show that they also satisfy (2.9). We say that $A \in M_n$ is a *generalized z-circulant* (where $z \in C$ has modulus 1) if there are complex numbers $\alpha_0, \ldots, \alpha_{n-1}$ such that

$$A = \sum_{i=0}^{n-1} \alpha_i C^i, \tag{2.12}$$

where C is given by

$$c_{ij} = \begin{cases} 1 & i = 1, \ldots, n-1, \quad j = i+1, \\ z & i = n, \quad j = 1, \\ 0 & \text{otherwise}. \end{cases} \tag{2.13}$$

We do not use the simpler notation "g-circulant" because it already has an accepted meaning. A generalized 1-circulant is what is usually referred to as a circulant. We are most interested in generalized 1-circulants in this paper. The matrix C defined in (2.13) is normal if and only if $|z| = 1$. It will be seen in the proof of Theorem 2.6 that the normality of C is essential.

The eigenvalues of a generalized z-circulant are very easy to find. Let $z = e^{i\theta}$ and let A be the generalized z-circulant

$$A = \sum_{j=0}^{n-1} \alpha_j C^j.$$

Then, since A is a polynomial in C, the eigenvalues of A are

$$\lambda_k = \sum_{j=0}^{n-1} \alpha_j \omega_k^j \quad k = 1, \ldots, n,$$

where $\omega_k = \exp((\theta + 2k\pi)/n), k = 1, \ldots, n$ are the eigenvalues of C. It will be seen in the next section that one can sometimes further simplify the expression for the λ_k.

Notice that the proof of the next result, which is one of our main results, depends only on the "if" part of Lemma 2.4.

THEOREM 2.6. *Let $A \in M_n$ be a generalized z-circulant. Then*

$$\|A\|_\infty = \frac{1}{n}\|A\|_1 = \frac{1}{n}\sum_{i=1}^{n} |\lambda_i(A)|, \tag{2.14}$$

where $\lambda_i(A)$ $i = 1, \ldots, n$ are the eigenvalues of A. Furthermore, there is a unitary generalized z-circulant U such that $|\!|\!|A|\!|\!|_\infty = \|A \circ U\|_\infty$.

Proof. Let $A \in M_n$ be a generalized z-circulant and let C be the matrix given by (2.13). The second equality in (2.14) follows from the fact that because C is normal so is A, hence the singular values of A are the absolute values of its eigenvalues. We show that A satisfies the condition (2.9) and that the first equality in (2.14) holds. Because A is normal $|A| = (A^*A)^{1/2}$ is a polynomial in A (in particular, $|A| = p(A)$, where p is a polynomial such that $p(\lambda_i) = |\lambda_i|$ for each eigenvalue λ_i of A), hence a polynomial in C. Thus $|A|$ has constant main diagonal entries. To prove the final statement note that because A is normal there is a unitary V that is a polynomial in A (hence in C) such that $VA = |A|$. Take $U = V^T$. □

Recently there has been some interest in the norm

$$|\!|\!|A|\!|\!|_\omega \equiv \max\{\omega(A \circ B) : \omega(B) \leq 1\},$$

where $\omega(A) \equiv \max\{|x^*Ax| : x \in C^n, x^*x = 1\}$ is the numerical radius of A. By an analogue of Lemma 2.2 ([2] or [13, Thm. 3.2]) one can derive $|\!|\!|A|\!|\!|_\omega = \|A\|_1/n$ (assuming A is a generalized z-circulant) from Theorem 2.6. Using this fact one can generalize all but one of the results in this paper to the numerical radius. The exception is Corollary 4.3, where it is necessary to use the fact that $\|UAV\|_\infty = \|A\|_\infty$ for all unitaries U and V; it is not sufficient to use merely $\|UAU^*\|_\infty = \|A\|_\infty$ for all unitaries U.

In the next lemma we identify a class of matrices A for which

$$\max\{\|A \circ B\|_\infty : \|B\|_\infty \leq 1\}$$

is attained at a matrix B that is essentially unique.

LEMMA 2.7. *Let $A \in M_n$. Assume that A is nonsingular, $|\!|\!|A|\!|\!|_\infty = \|A\|_1/n$, and that $U \equiv A|A|^{-1}$ is irreducible. Let $V \in M_n$ and suppose that $\|V\|_\infty \leq 1$. Then*

$$\|A \circ V\|_\infty = |\!|\!|A|\!|\!|_\infty = \|A\|_1/n \tag{2.15}$$

if and only if

$$V = W_1\bar{U}W_2 \quad \textit{where} \quad W_1, W_2 \in M_n \textit{ are diagonal unitary.} \tag{2.16}$$

Proof. The "if" part of the lemma is clear from Lemma 2.4.

Let us consider the "only if" part. Let A satisfy the conditions of the lemma and let $V \in M_n$ be a contraction satisfying (2.15). We must show that V is of the form (2.16). Then

$$B \equiv \begin{pmatrix} |A^*| & A \\ A^* & |A| \end{pmatrix} \geq 0, \quad C \equiv \begin{pmatrix} I & -V \\ -V^* & I \end{pmatrix} \geq 0.$$

With these definitions

$$B \circ C = \begin{pmatrix} (\mathrm{tr}|A|/n)I & -A \circ V \\ -(A \circ V)^* & (\mathrm{tr}|A|/n)I \end{pmatrix} \geq 0.$$

By assumption $\|A \circ V\|_\infty = \mathrm{tr}|A|/n$ so $B \circ C$ must be singular. The singularity of $B \circ C$ (and the fact that $B \geq 0$, $C \geq 0$) imply that there is a nonzero diagonal matrix $D \in M_{2n}$ such that $BD^*C^TD = 0$ [13, Lemma 1.2]. Let us first assume that D is entrywise nonnegative. Let $D_1, D_2 \in M_n$ be such that $D = D_1 \oplus D_2$. Writing $B(D^*C^TD) = 0$ in block form we have

$$\begin{pmatrix} |A^*| & A \\ A^* & |A| \end{pmatrix} \begin{pmatrix} D_1^2 & -D_1\bar{V}D_2 \\ -D_2V^TD_1 & D_2^2 \end{pmatrix} = 0.$$

Considering the 1,1 block of the product we have

$$|A^*|D_1^2 = AD_2V^TD_1 = |A^*|UD_2V^TD_1,$$

which (using the nonsingularity of A) gives

$$U^*D_1^2 = D_2V^TD_1$$

or equivalently

$$D_1^2U = D_1\bar{V}D_2. \tag{2.17}$$

Similarly, consideration of the 2, 2 entry gives

$$UD_2^2 = D_1\bar{V}D_2. \tag{2.18}$$

Combining (2.17) and (2.18) we have $D_1^2U = UD_2^2$. Thus $(D_1^2)_{ii} = (D_2^2)_{jj}$ for any i, j for which $u_{ij} \neq 0$. The irreducibility of U now ensures that $D_1^2 = D_2^2 = c^2I$ for some $c > 0$, and (because D_1 and D_2 are entrywise nonnegative) $D_1 = D_2 = cI$. Thus $\bar{V} = U$ or equivalently $V = \bar{U}$.

Now let us consider the case where the diagonal matrix D is not nonnegative. In this case let $D = (W_1D_1) \oplus (W_2D_2)$ where $D_1, D_2, W_1, W_2 \in M_n$ are diagonal, W_1, W_2 are unitary and D_1, D_2 are nonnegative and proceed as before. □

3. Some special matrices. In this section we compute or bound $|||A|||_\infty$ for some matrices related to circulants. The following constants will be useful.

$$\gamma_n \equiv \frac{1}{n}\sum_{j=1}^{n} |\cot(2j-1)\pi/2n| = \frac{2}{n}\sum_{j=1}^{[n/2]} \cot(2j-1)\pi/2n,$$

$$\hat{\gamma}_n \equiv \frac{1}{n}\sum_{j=1}^{n} |\csc(2j-1)\pi/2n|.$$

Since $|\cot\theta| < |\csc\theta| < |\cot\theta| + 1$ for $\theta \neq k\pi/2$, it follows that $\gamma_n < \hat{\gamma}_n < \gamma_n + 1$. Approximating the sums by integrals one can show that $\hat{\gamma}_n - \gamma_n \to (1/\pi)\log 2$, $\gamma_n/\log n \to 2/\pi$, and $\hat{\gamma}_n/\log n \to 2/\pi$, as $n \to \infty$. The sequences γ_n and $\hat{\gamma}_n$ are strictly increasing (as will be immediate from the proof of the next lemma).

LEMMA 3.1. *For $i = 1, 2, \ldots$*

$$|||S_n|||_\infty = \gamma_n \quad \textit{and} \quad |||\hat{S}_n|||_\infty = \hat{\gamma}_n. \tag{3.1}$$

Proof. We first prove the result for $\hat{S}_n$. Let C be the matrix given by (2.13) when $z = -1$. Then $\hat{S}_n = I + C + C^2 + \cdots + C^{n-1}$. The eigenvalues of C are

$$\omega_k = e^{i(2k-1)\pi/n}, \quad k = 1, \ldots, n$$

from which it follows that the eigenvalues of $\hat{S}_n$ are $\lambda_k = (1 - \omega_k^n)/(1 - \omega_k)$, which using half-angle formulae gives

$$\lambda_k = 1 + i \ \cot(2k-1)\pi/2n.$$

One can show that $|\lambda_k| = |\sin(2k-1)\pi/2n|^{-1}$. The result for $\hat{S}_n$ follows from this.

To see that $|||S_n|||_\infty = \gamma_n$, just note that $S_n = \hat{S}_n - I$ and, hence, that its eigenvalues are $i\cot(2k-1)\pi/2n$. □

Pokrzywa [15] has shown that

$$\max\{\|Z - Z^*\|_\infty : Z \in M_n, \|Z + Z^*\|_\infty \le 1, Z \text{ has real eigenvalues}\} = \gamma_n. \tag{3.2}$$

This fact is essentially the same as $|||S_n|||_\infty = \gamma_n$ as we will demonstrate. (This result was originally proved to help bound the eigenvalues of an arbitrary perturbation of a Hermitian matrix; see the discussion at the end of §4.) To show this we first prove that

$$\begin{aligned}\max\{\|Z - Z^*\|_\infty &: Z \in M_n, \|Z + Z^*\|_\infty \le 1, Z \text{ has real eigenvalues}\}\\ &= \max\{\|S_n \circ H\|_\infty : H \in M_n, \|H\|_\infty \le 1, H = H^*\}.\end{aligned} \tag{3.3}$$

Let $Z \in M_n$ have real eigenvalues then, by Schur's Upper Triangularization Theorem, there is a unitary $V \in M_n$ such that VZV^* is upper triangular. Let $VZV^* = D + U$, where D is a real diagonal matrix (real because Z has real eigenvalues) and U is strictly upper triangular. Then, since $S_n \circ D = 0$, $S_n \circ U = U$, and $S_n \circ U^* = -U^*$, we have

$$VZV^* - (VZV^*)^* = U - U^* = S_n \circ (2D + U + U^*) = S_n \circ [VZV^* + (VZV^*)^*].$$

Thus, by the unitary invariance of $\|\cdot\|_\infty$ the left-hand side of (3.3) is no greater than the right-hand side. Conversely, given $H \in H_n$ let U be strictly upper triangular and let D be real and diagonal such that $H = U^* + 2D + U$. Let $Z = D + U$. Then $H = Z + Z^*$ and $S_n \circ H = U - U^* = Z - Z^*$. This gives the reverse inequality.

Because S_n is skew-Hermitian (hence iS_n is Hermitian) Corollary 3.3 in [13] says that (3.3) is equal to

$$\max\{\|S_n \circ H\|_\infty : H \in M_n, \|H\|_\infty \le 1\} = |||S_n|||_\infty = \gamma_n.$$

Now we will turn our attention to the matrix $[(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)]$.

THEOREM 3.2. *Let $\lambda_1, \ldots, \lambda_n$ be positive real numbers. Then*

$$|||[\,|(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)|\,]|||_\infty \le 2. \tag{3.4}$$

Furthermore, 2 is the best possible bound that is independent of n.

Proof. First we will show $|||[\,|(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)|\,]|||_\infty \le 2$. Note that

$$|(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)| = 1 - 2\min\{\lambda_i, \lambda_j\}/(\lambda_i + \lambda_j).$$

Thus

$$[|(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)|] = J_n - [2\min\{\lambda_i, \lambda_j\}/(\lambda_i + \lambda_j)]. \tag{3.5}$$

The Cauchy matrix $[(\lambda_i + \lambda_j)^{-1}]$ is positive semidefinite (see e.g., [9, §7.1, problem 17]). The matrix $[2\min\{\lambda_i, \lambda_j\}]$ is positive semidefinite. Assume without loss of generality that $\lambda_1 \geq \cdots \geq \lambda_n$, then

$$2[\min\{\lambda_i, \lambda_j\}] = 2\sum_{i=1}^{n}(\lambda_i - \lambda_{i+1})J_i \oplus 0_{n-i},$$

where $\lambda_{n+1} = 0$. Thus, by the Schur Product Theorem the second matrix in (3.5) is positive semidefinite. So, using (1.2) for the last inequality, we have

$$|||\,[\,|(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)|\,]\,|||_\infty \leq |||J_n|||_\infty + |||\,[2\min\{\lambda_i, \lambda_j\}/(\lambda_i + \lambda_j)]\,|||_\infty \leq 1 + 1 = 2.$$

Now we will show that 2 is the best bound in (3.4) that is independent of n. Note that

$$\lim_{m\to\infty}[|(m^i - m^j)/(m^i + m^j)|] = J_n - I.$$

$J_n - I$ is a circulant with one eigenvalue $n-1$ and $n-1$ eigenvalues -1 so

$$|||J_n - I|||_\infty = 2(n-1)/n.$$

(Bhatia, Choi, and Davis [6, Thm. 2.1] have also shown this.) Thus 2 is the best bound independent of n. □

COROLLARY 3.3. *Let*

$$\beta_n = \sup\{|||[(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)]|||_\infty : \lambda \in R_+^n\}.$$

Then

$$\gamma_n \leq \beta_n \leq 2\gamma_n.$$

Proof. To get the lower bound note that S_n is the limit of matrices of the form $[(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)]$ (see the proof of Theorem 3.2). For the upper bound use the fact that we may assume without loss of generality that $\lambda_1 \geq \cdots \geq \lambda_n$, in which case $[(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)] = S_n \circ [\,|(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)|\,]$. □

Notice that the upper and lower bounds in Corollary 3.3 differ by a factor of 2. One would like to tighten these bounds since there are many bounds in the rest of the paper that involve β_n.

COROLLARY 3.4. *Let $m \leq n$ and define*

$$\delta_{m,n} \equiv \sup\{|||[(\lambda_i - \mu_j)/(\lambda_i + \mu_j)]|||_\infty : \lambda \in R_+^m, \mu \in R_+^n\}. \tag{3.6}$$

Then

$$\gamma_m \leq \beta_m \leq \delta_{m,n} \leq \beta_{m+n} \leq 2\gamma_{m+n}. \tag{3.7}$$

Proof. Let

$$D_1 = \left[\frac{\lambda_i - \lambda_j}{\lambda_i + \lambda_j}\right]_{i,j=1}^{m}, \quad D_2 = \left[\frac{\lambda_i - \mu_j}{\lambda_i + \mu_j}\right]_{i=1,j=1}^{m,\ n}, \quad \text{and} \quad D_3 = \left[\frac{\mu_i - \mu_j}{\mu_i + \mu_j}\right]_{i,j=1}^{n}.$$

Since D_2 is a submatrix of

$$D \equiv \begin{pmatrix} D_1 & D_2 \\ -D_2^T & D_3 \end{pmatrix}$$

it follows that $|||D_2|||_\infty \leq |||D|||_\infty \leq \beta_{m+n}$.

For the lower bound use the fact if we take $\mu_i = \lambda_i$ for $i = 1, \ldots, m$ then

$$\left[\frac{\lambda_i - \lambda_j}{\lambda_i + \lambda_j}\right]_{i,j=1}^{m}$$

is a submatrix of

$$\left[\frac{\lambda_i - \mu_j}{\lambda_i + \mu_j}\right]_{i=1,j=1}^{m,\ n}. \qquad \square$$

Davies [7, Prop. 4] has proved the following bounds that are independent of n for the Hadamard operator norm of S_n with respect to the Schatten p-norms:

$$|||[(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)]_{i,j=1}^{n}|||_p \leq \begin{cases} cp & 2 \leq p < \infty, \\ cp/(1-p) & 1 < p \leq 2, \end{cases}$$

where c is an absolute constant (independent of p and n).

Now we will turn our attention to the norm of triangular truncation.

COROLLARY 3.5. *For any* $n = 1, 2 \ldots$

$$\frac{(n+1)\hat{\gamma}_n}{2n} \leq |||\hat{T}_n|||_\infty \leq \frac{\hat{\gamma}_n + 1}{2}. \tag{3.8}$$

Proof. For the upper bound use $\hat{T}_n = (\hat{S}_n + J_n)/2$. So

$$|||\hat{T}_n|||_\infty \leq (1/2)[|||\hat{S}_n|||_\infty + |||J_n|||_\infty] = (1/2)[\hat{\gamma}_n + 1].$$

For the lower bound let $U = |\hat{S}_n|^{-1}\hat{S}_n$ (from the proof of Lemma 3.1 we can see that $\hat{S}_n$ is nonsingular). Then $U\hat{S}_n = \hat{S}_n U = |\hat{S}_n|$ since $\hat{S}_n$ is normal. So we know (from the proof of Lemma 2.4) that

$$(e_n/\sqrt{n})^*(\hat{S}_n \circ U^T)(e_n/\sqrt{n}) = \|\hat{S}\|_1/n = \hat{\gamma}_n.$$

Thus,

$$\begin{aligned} |||\hat{T}_n|||_\infty &\geq \|\hat{T}_n \circ U^T\|_\infty \\ &\geq (1/2)(e_n/\sqrt{n})^*(\hat{S}_n \circ U^T + J_n \circ U^T)(e_n/\sqrt{n}) \\ &= (1/2)[\hat{\gamma}_n + (1/n)e_n^* U^T e_n]. \end{aligned}$$

We will have proved the lower bound if we can show that $e_n^* U^T e_n = \hat{\gamma}_n$. Because U is real we have $e_n^* U^T e_n = e_n^* U e_n = (e_n^* U)e_n$. The first row of $\hat{S}_n$ is e_n^* and so $e_n^* U$ is the first row of $\hat{S}_n U = |\hat{S}_n|$. Thus $e_n^* U e_n$ is the sum of the entries in the first row of $|\hat{S}_n|$. We know that

$$|\hat{S}_n| = \sum_{i=0}^{n-1} \alpha_i C^i$$

for some real coefficients α_i (where C is the matrix given by (2.13) with $z = -1$). Since $|\hat{S}_n|$ is positive definite and real it must be symmetric. The symmetry of $\hat{S}_n$ requires that $\alpha_i = -\alpha_{n-i}$, $i = 1, 2, \ldots, n-1$. Thus the sum of the first row of $|\hat{S}_n|$ is

$$\sum_{i=0}^{n-1} \alpha_i = \alpha_0 = |\hat{S}_n|_{11} = \mathrm{tr}|\hat{S}_n|/n = \hat{\gamma}_n.$$

(We have used the fact that $|\hat{S}_n|$ has constant main diagonal entries for the third equality.) □

We can obtain a slightly weaker lower bound on $|||\hat{T}_n|||_\infty$ with considerably less effort. Since $\hat{S}_n = \hat{T}_n - T_n^*$ and $|||T_n|||_\infty = |||\hat{T}_{n-1}|||_\infty \leq |||\hat{T}_n|||_\infty$ we have

$$\hat{\gamma}_n = |||\hat{S}_n|||_\infty \leq |||\hat{T}_n|||_\infty + |||T_n^*|||_\infty \leq 2|||\hat{T}_n|||_\infty.$$

Angelos, Cowen, and Narayan [3] have determined $|||\hat{T}_n|||_\infty$ in closed form for $n = 2, 3, 4$ and bounded $|||\hat{T}_n|||_\infty$ for general n. One can show that our bounds are better than those in [3, Thm. 1].

We have only considered the norms $|||\cdot|||_\infty$ and $|||\cdot|||_1$. Using the fact that for any other unitarily invariant norm $||\cdot||$ on M_n the corresponding Hadamard operator norm satisfies

$$|||A||| \equiv \max\{||A \circ B|| : ||B|| \leq 1, B \in M_n\} \leq |||A|||_\infty, \tag{3.9}$$

we have upper bounds on $|||A|||$. These bounds may not be very good (cf. the results of Davies mentioned earlier in this section).

4. Applications. In this section we apply the results of the previous section to problems involving commutators, the absolute value, and the eigenvalues of an arbitrary perturbation of a Hermitian matrix. We recast these problems in terms of bounding the Hadamard operator norm of an appropriate matrix. One advantage of this approach is that, in view of Lemma 2.1, it is immediate that the same inequality holds for the spectral norm and the trace norm, and by (3.9) our upper bounds are valid for all unitarily invariant norms. In the case of the Schatten p-norms ($1 < p < \infty$) Davies [7] (see his results quoted in the previous section) has obtained stronger results (that are independent of the dimension n). The bound on the norm of $|||T|||$ is useful in perturbation bounds for triangular factorizations matrices. For example, it is used in [8] for a perturbation bound on the Cholesky factor of a positive definite matrix. We do not discuss these applications here.

First we obtain bounds on

$$c_n \equiv \sup\{||\, |A| - |B|\, ||_\infty : ||A - B||_\infty = 1,\ \ A, B \in M_n\}. \tag{4.1}$$

In [12] Kato showed that c_n does indeed depend on n. (If one looks carefully at his proof one sees that he showed that $c_n \geq (1/3)(\log n)^{1/2} - 1$.) Davies [7, Thms. 13, 14] showed that there are positive constants k_1, k_2 such that

$$k_1 \log n \leq c_n \leq k_2 \log n.$$

(Actually, he defined c_n using $||\cdot||_1$ rather than $||\cdot||_\infty$ in (4.1); but using Lemma 4.1, Lemma 4.2, and the duality result in Lemma 2.1, one can show that defining c_n with the norm $||\cdot||_1$ rather than $||\cdot||_\infty$ changes c_n by at most 1.)

LEMMA 4.1. *For any* $n = 1, 2, \ldots$

$$c_n = \sup\left\{ \left\| \frac{d}{dt}|A + tC|_{t=0} \right\|_\infty : A, C \in M_n, \|C\|_\infty \leq 1, A \text{ nonsingular} \right\}. \tag{4.2}$$

Proof. One can show that the left-hand side of (4.2) is greater than or equal to the right-hand side by using the definition of the derivative. In order to show the reverse inequality it is sufficient to show that

$$\|\, |A| - |B| \,\|_\infty \leq \hat{c}_n \|A - B\|_\infty, \tag{4.3}$$

where

$$\hat{c}_n = \sup\left\{ \left\| \frac{d}{dt}|A + tC|_{t=0} \right\|_\infty : A, C \in M_n, \|C\|_\infty \leq 1, A \text{ nonsingular} \right\}.$$

We will first prove this in the special case where $A + t(B - A)$ is nonsingular for all $t \in [0, 1]$. In this case, $f(t) = |A + t(B - A)|$ is differentiable for all $t \in [0, 1]$, so

$$\begin{aligned} \|\, |A| - |B| \,\|_\infty &= \left\| \int_0^1 f'(t)dt \right\|_\infty \\ &\leq \int_0^1 \|f'(t)\|_\infty dt \\ &\leq \int_0^1 \hat{c}_n \|A - B\|_\infty dt = \hat{c}_n \|A - B\|_\infty. \end{aligned}$$

Now let us consider the general case. It is known that for any $A, B \in M_n$

$$\|\, |A| - |B| \,\|_\infty \leq \|\, |A| - |B| \,\|_2 \leq \sqrt{2} \|\, A - B \,\|_2 \leq \sqrt{2n} \|\, A - B \,\|_\infty.$$

(See [7, Lemma 2] for the second inequality.)

Take any $A, B \in M_n$ and any $\epsilon > 0$. Choose $s > 0$ such that $s < \epsilon/\sqrt{2n}$ and both $A + sI$ and $B + sI$ are nonsingular. Define $A(t) = A + sI + t(B - A)$. Then $A(t)$ is singular at at most n values of t in $[0, 1]$. Let $0 < t_1 < \cdots < t_k < 1$ be the points in $[0, 1]$ at which $A(t)$ is singular. Set $t_0 = 0$ and $t_{k+1} = 1$. Choose $\eta > 0$ such that $2\eta < t_i - t_{i-1}, i = 1, \ldots, k + 1$ and $\eta < \epsilon/\sqrt{2n}$. Then

$$\begin{aligned} \|\, |A(t_i)| - |A(t_{i-1})| \,\|_\infty &\leq \|\, |A(t_i)| - |A(t_i - \eta)| \,\|_\infty + \|\, |A(t_i - \eta)| - |A(t_{i-1} + \eta)| \,\|_\infty \\ &\quad + \|\, |A(t_{i-1} + \eta)| - |A(t_{i-1})| \,\|_\infty \\ &\leq \sqrt{2n} \|\eta(A - B)\|_\infty + \hat{c}_n \|(t_i - t_{i-1})(A - B)\|_\infty \\ &\quad + \sqrt{2n} \|\eta(A - B)\|_\infty \\ &\leq [(t_i - t_{i-1})\hat{c}_n + 2\epsilon] \|A - B\|_\infty. \end{aligned}$$

So now

$$\begin{aligned} \|\, |A| - |B| \,\|_\infty &\leq \|\, |A| - |A(0)| \,\|_\infty + \sum_{i=1}^{k+1} \|\, |A(t_i)| - |A(t_{i-1})| \,\|_\infty \\ &\quad + \|\, |A(1)| - |B| \,\|_\infty \\ &\leq \hat{c}_n \|A - B\|_\infty + \epsilon[2(k + 1)\|A - B\|_\infty + 2]. \end{aligned}$$

Let $\epsilon \downarrow 0$ to get (4.3). □

The next lemma allows us to transform the problem of determining c_n into one involving Hadamard products.

LEMMA 4.2. *Let $A, C \in M_n$ be given with A nonsingular. Let $\sigma_i, i = 1, \ldots, n$ be the singular values of A and let Σ be a diagonal matrix with the singular values of A on the diagonal, and let $A = U\Sigma V$ be a singular value decomposition of A. Let $f(t) = |A + tC|$. Then*

$$f'(0) = V^*(H + [(\sigma_i - \sigma_j)/(\sigma_i + \sigma_j)] \circ K)V, \tag{4.4}$$

where

$$2H = (U^*CV) + (U^*CV)^* \quad and \quad 2K = (U^*CV) - (U^*CV)^*. \tag{4.5}$$

Proof. Generalize (2.18a) in [4] to the complex case. Alternatively, apply [10, Thm. 6.6.30] to the function $f(t) = [(A + tC)^*(A + tC)]^{1/2}$. □

COROLLARY 4.3. *Let $n = 2, 3, \ldots$. Then*

$$\gamma_n \leq \beta_n \leq c_n \leq 1 + \beta_n \leq 2\gamma_n + 1.$$

Proof. To prove the lower bound on c_n, let $\sigma^{(k)} \in R^n_+$ be such that

$$\left|\!\left|\!\left| \left[\frac{\sigma_i^{(k)} - \sigma_j^{(k)}}{\sigma_i^{(k)} + \sigma_j^{(k)}} \right] \right|\!\right|\!\right|_\infty \to \beta_n,$$

and let $W_k \in M_n$ be skew-Hermitian contractions such that

$$\left|\!\left|\!\left| \left[\frac{\sigma_i^{(k)} - \sigma_j^{(k)}}{\sigma_i^{(k)} + \sigma_j^{(k)}} \right] \right|\!\right|\!\right|_\infty = \left\| \left[\frac{\sigma_i^{(k)} - \sigma_j^{(k)}}{\sigma_i^{(k)} + \sigma_j^{(k)}} \right] \circ W_k \right\|_\infty .$$

Corollary 3.3 in [13] guarantees that W_k can be chosen to be skew-Hermitian since $[\,(\sigma_i^{(k)} - \sigma_j^{(k)})/(\sigma_i^{(k)} + \sigma_j^{(k)})\,]$ is skew-Hermitian. Then

$$\lim_{k \to \infty} \left\| \frac{d}{dt} |\mathrm{diag}(\sigma_1^{(k)}, \ldots, \sigma_n^{(k)}) + tW_k|_{t=0} \right\|_\infty = \beta_n.$$

Thus, by Lemma 4.2, it follows that $c_n \geq \beta_n$.

Now we prove the upper bound on c_n, again using Lemma 4.1. Take $A, C \in M_n$ with A nonsingular and $\|C\|_\infty \leq 1$. Let $A = U\Sigma V$ be a singular value decomposition of A. Let H and K be given by (4.5), and therefore also contractions. Then, by Lemma 4.1

$$\begin{aligned}
\left\| \frac{d}{dt} |A + tC|_{t=0} \right\|_\infty &= \|V^*(H + [(\sigma_i - \sigma_j)/(\sigma_i + \sigma_j)] \circ K)V\|_\infty \\
&= \|H + [(\sigma_i - \sigma_j)/(\sigma_i + \sigma_j)] \circ K\|_\infty \\
&\leq \|H\|_\infty + \|[(\sigma_i - \sigma_j)/(\sigma_i + \sigma_j)] \circ K\|_\infty \\
&\leq 1 + |\!|\!|[(\sigma_i - \sigma_j)/(\sigma_i + \sigma_j)]|\!|\!|_\infty \\
&\leq 1 + \beta_n. \quad \square
\end{aligned}$$

THEOREM 4.4. *Let $A, B \in M_n$ be given with A Hermitian. Then*

$$\|\, |A|B - B|A| \,\| \leq (1 + \beta_n)\|AB - BA\|$$

for all unitarily invariant norms $\|\cdot\|$. *If* $\|\cdot\|$ *is* $\|\cdot\|_\infty$ *or* $\|\cdot\|_1$, *then given* $n \geq 2$ *and* $\epsilon > 0$, *there are* $A, B \in M_n$ *with* A *Hermitian such that*

$$\|\,|A|B - B|A|\,\| \geq (\beta_{[n/2]} - \epsilon)\|AB - BA\|.$$

Proof. Without loss of generality we may assume that A is diagonal and nonsingular. Let $\lambda_i, i = 1, \ldots, k$ be the positive eigenvalues of A and let $-\mu_{k+i}, i = 1, \ldots, n-k$ be the negative eigenvalues of A. Then one can check that

$$|A|B - B|A| = C \circ (AB - BA),$$

where

$$C = \begin{pmatrix} J_k & D \\ -D & -J_{n-k} \end{pmatrix}$$

and D is the $k \times (n-k)$ matrix with i, j entry $(\lambda_i - \mu_j)/(\lambda_i + \mu_j)$. So to obtain the inequality we need only find a suitable bound on $|\!|\!|C|\!|\!|_\infty$. Using (3.9) and Corollary 3.4 we have

$$\begin{aligned} |\!|\!|C|\!|\!| \leq |\!|\!|C|\!|\!|_\infty &= \left|\!\left|\!\left| \begin{pmatrix} J_k & D \\ -D^* & -J_{n-k} \end{pmatrix} \right|\!\right|\!\right|_\infty \\ &\leq \left|\!\left|\!\left| \begin{pmatrix} J_k & 0 \\ 0 & -J_{n-k} \end{pmatrix} \right|\!\right|\!\right|_\infty + \left|\!\left|\!\left| \begin{pmatrix} 0 & D \\ -D^* & 0 \end{pmatrix} \right|\!\right|\!\right|_\infty \\ &\leq 1 + \beta_n. \end{aligned}$$

To prove the lower bound on $\|\,|A|B - B|A|\,\|$ use the usual limiting argument. □

One can prove a variety of other results involving commutators and the absolute value by reducing them to problems involving the Hadamard product with a matrix of the form $[(\lambda_i - \lambda_j)/(\lambda_i + \lambda_j)]$.

Finally, we apply our results to the problem of bounding the change in the eigenvalues of a Hermitian matrix subjected to an arbitrary perturbation. The bound (4.6) is similar to one given by Kahan (see [5, §23] for the details) and is based on (3.2). A strengthening of this result in the case where the eigenvalues of A are clustered is given in [16]. The lower bound (4.7) is new and improves on that given in [5, Ex. 23.4] by a factor of $\pi/2$.

THEOREM 4.5. *Let* $A, X \in M_n$ *be given with* A *Hermitian and* X *arbitrary. Let* $\alpha_1 \geq \cdots \geq \alpha_n$ *be the eigenvalues of* A *and let* $\lambda_1, \ldots, \lambda_n$ *be the eigenvalues of* $A + X$ *ordered such that* $\mathrm{Re}\,\lambda_1 \geq \cdots \geq \mathrm{Re}\,\lambda_n$. *Then*

$$\max_{1 \leq i \leq n} |\alpha_i - \lambda_i| \leq (\gamma_n + 2)\|X\|_\infty. \tag{4.6}$$

Furthermore, for any given positive integer n *there are* $A, X \in M_n$ *with* A *Hermitian and* $X \neq 0$ *such that no matter how* α_i, λ_i $i = 1, \ldots, n$, *the eigenvalues of* A *and* $A + X$, *respectively, are ordered*

$$\max_{1 \leq i \leq n} |\alpha_i - \lambda_i| \geq \gamma_n\|X\|_\infty. \tag{4.7}$$

Proof. See [5, §23] for the proof of (4.6). We prove (4.7). Let U be a real unitary matrix such that $|S_n| = US_n$. Then since S_n is normal we also have $|S_n| = S_nU$ and hence $|S_n| = (S_nU)^* = U^*S_n^* = -U^*S_n$. Thus $(1/2)(U - U^*)S_n = |S_n|$. Let

$X = (1/2)(U - U^*)^T = (1/2)(U^T - U)$ and let $A = S_n \circ X$. Clearly X is a contraction, so by the final statement in Lemma 2.4 $\|A\|_\infty = \|S_n \circ X\|_\infty = \gamma_n$. Because A is Hermitian it follows that one of the eigenvalues of A must be $\pm\gamma_n$. Because X is real and skew-Hermitian its main diagonal entries are 0 and so $A + X = S_n \circ X + X$ is strictly upper triangular and so all its eigenvalues are 0. Thus we have (4.7). □

We could also prove (4.7) by using the matrix Z that attains the maximum in (3.2). The advantage of our approach is that with a little work it is possible to find explicitly the unitary matrix U in the proof and thus find A and X explicitly.

5. A counterexample.

In this section we show that the following conjecture, which is slightly weaker than Conjecture 6.8 in [14], is false.

CONJECTURE 5.1. *Let $A \in M_n$ have real entries and suppose that there is a real unitary $U \in M_n$ such that all the entries of $A \circ U$ are positive then*

$$\| A \|_\infty = \max\{\|A \circ B\|_\infty : \|B\|_\infty \le 1, a_{ij}b_{ij} \ge 0, i, j = 1, \ldots, n, B \in M_n(R)\}. \tag{5.1}$$

Let

$$A = \begin{pmatrix} 10 & 1 & -2 \\ 2 & 10 & 1 \\ -1 & 2 & 10 \end{pmatrix} \quad \text{and} \quad U_1 = \begin{pmatrix} 1/3 & 2/3 & -2/3 \\ 2/3 & 1/3 & 2/3 \\ -2/3 & 2/3 & 1/3 \end{pmatrix}.$$

Then U_1 is a unitary matrix and the entries of $A \circ U$ are positive. Because A is a generalized z-circulant $\| A \|_\infty = \|A\|_1/3$. Also A is nonsingular. Let $U \equiv A|A|^{-1}$. The U is irreducible. One can numerically check that

$$U \approx \begin{pmatrix} .9981 & -.0443 & -.0424 \\ .0424 & .9981 & -.0443 \\ .0443 & .0424 & .9981 \end{pmatrix}.$$

By Lemma 2.7, if $V \in M_n$ is a contraction such that $\|A \circ V\|_\infty = \| A \|_\infty$ then $V = W_1\bar{U}W_2$ for some diagonal unitaries W_i. However, one can show that there is no way to choose diagonal unitaries W_i such that $A \circ (W_1\bar{U}W_2) = W_1(A \circ \bar{U})W_2$ is entrywise nonnegative. Thus Conjecture 5.1 is false.

REFERENCES

[1] T. ANDO, *On the arithmetic-geometric-harmonic-mean inequality for positive definite matrices*, Linear Algebra Appl., 52/53 (1983), pp. 31–37.

[2] T. ANDO AND K. OKUBO, *Induced norms of the Schur multiplier operator*, Linear Algebra Appl., 147 (1991), pp. 181–199.

[3] J. ANGELOS, C. COWEN, AND S. K. NARAYAN, *Triangular truncation and finding the norm of a Hadamard multiplier*, Linear Algebra Appl., 170 (1992), pp. 117–135.

[4] A. BARRLUND, *Perturbation bounds on the polar decomposition*, BIT, 30 (1989), pp. 101–113.

[5] R. BHATIA, *Perturbation Bounds for Matrix Eigenvalues*, Pitman Research Notes in Mathematics 162, Longman Scientific and Technical, New York, 1987.

[6] R. BHATIA, M.-D. CHOI, AND C. DAVIS, *Comparing a matrix to its off-diagonal part*, Oper. Theory: Adv. Appl., 40 (1989), pp. 151–164.

[7] E. B. DAVIES, *Lipschitz continuity of functions of operators in the Schatten classes*, J. London Math. Soc., 37 (1988), pp. 148–157.

[8] Z. DRMAC, M. OMLADIC, AND K. VESELIC, *On the perturbation of the Cholesky factorization*, preprint, 1992.

[9] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[10] ———, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[11] R. A. HORN AND R. MATHIAS, *Cauchy-Schwarz inequalities associated with positive semidefinite matrices*, Linear Algebra Appl., 142 (1990), pp. 63–82.

[12] T. KATO, *Continuity of the map $S \rightarrow |S|$ for linear operators*, Proc. Japan Acad., 49 (1973), pp. 157–160.

[13] R. MATHIAS, *Matrix completions, Hadamard products and norms*, Proc. Amer. Math. Soc., to appear.

[14] R. MCEACHIN, *Analysis of an Inequality Concerning Perturbation of Self-Adjoint Operators*, Ph.D. thesis, Mathematics Dept., Univ. of Illinois at Urbana-Champaign, Urbana, IL, 1990.

[15] A. POKRZYWA, *Spectra of operators with fixed imaginary part*, Proc. Amer. Math. Soc., 81 (1981), pp. 359–364.

[16] A. SCHÖNHAGE, *Arbitrary perturbations of Hermitian matrices*, Linear Algebra Appl., 24 (1979), pp. 143–149.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, pp. 1168–1190, October 1993

021

A NON-INTERIOR-POINT CONTINUATION METHOD FOR LINEAR COMPLEMENTARITY PROBLEMS*

BINTONG CHEN† AND PATRICK T. HARKER‡

Abstract. This paper presents a continuation method for linear complementarity problems based on a new smooth equation formulation. In particular, the case of a linear complementarity problem defined by a positive semidefinite or P_0 matrix is studied in detail. Extensive numerical testing of the continuation method is performed for both problems in the literature and randomly generated problems.

Key words. linear complementarity, continuation, interior-point methods, P_0-matrix

AMS subject classification. 90C33

1. Introduction.

This paper presents a continuation method for linear complementarity problems (LCP), which is loosely based on notions from the class of interior-point algorithms for these problems. In particular, the current research was motivated by the path following algorithm by Kojima, Mizuno, and Yoshise [14]. Given a monotone LCP, the Kojima et al. algorithm creates a strictly feasible path, called the *path of centers*, by relaxing the complementarity condition while maintaining the feasibility condition. It was shown that by following the path properly, the algorithm will converge to a solution of the LCP. The algorithm is theoretically significant since its computational complexity is bounded by a polynomial function of the problem size. However, like all interior-point algorithms, the method has the following restrictions. First, by construction, the initial homotopy parameter is large and the starting point is normally far away from the solution of the problem unless a good estimate close to the path is available. Second, each new iterate is obtained by Newton's method at the current iterate, and all intermediate iterates are required to stay interior, which sometimes prevents the algorithm from using larger step sizes. It is clear that the main effort in the above algorithm is spent on tracking the *feasible* path, or the path of centers, and preventing each iterate from going out of the feasible region. The natural way to overcome this difficulty is to increase the attraction domain of the feasible path. The method proposed herein actually identifies the feasible path, which means, under appropriate assumptions, that the attraction domain of the feasible path is all of $\Re^n$ instead of the positive orthant $\Re^n_+$. Therefore, the algorithm can start at any initial point whether feasible or not and, in addition, each intermediate point does not have to stay feasible.

The basis of the research reported herein lies in a prior study on the general problem of solving monotone variational inequalities by an interior-pointlike continuation method [3]. Thus, we shall refer to this paper in a few of the proofs in order to keep the paper to a manageable length.

The remainder of this paper is structured as follows. The next section provides a brief overview of the related LCP literature. Section 3 presents the basic algorithm and related theoretical results. The rest of the paper is devoted to computational

* Received by the editors January 20, 1991; accepted for publication (in revised form) March 2, 1992.

† Department of Management and Systems, College of Business and Economics, Washington State University, Pullman, Washington 99164-4726 (chenbi@wsuvm1.csc.wsu.edu).

‡ Decision Sciences Department, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6366 (harker@wharton.upenn.edu).

tests of the proposed algorithm, and conclusions are presented in the last section.

1.1. Notation. The following notation will be used throughout the paper. Vectors are denoted by boldface lowercase Roman letters, such as $\mathbf{x}$. All vectors are column vectors, unless explicitly stated otherwise. Row vectors are the transpose of column vectors; for example, $\mathbf{x}^T$ denotes the row vector $(x_1, \ldots, x_n)$. For notational simplicity, $(\mathbf{x}_1^T, \mathbf{x}_2^T)^T$ is sometimes simplified as $(\mathbf{x}_1, \mathbf{x}_2)^T$. $\Re^n$, $\Re^n_+$, and $\Re^n_{++}$ denote, respectively, n-dimensional Euclidean space, the nonnegative orthant of $\Re^n$, and the strictly positive orthant of $\Re^n$. Matrices are denoted by boldface capital Roman letters, such as $\mathbf{M}$. The ith row of the $n \times n$ matrix $\mathbf{M}$ is denoted by $\mathbf{M}_{i\cdot}$, and the jth column is given as $\mathbf{M}_{\cdot j}$. Scalar-valued functions are denoted by Roman or Greek letters, such as f and θ. Vector-valued functions are denoted by boldface Roman or Greek letters, such as $\mathbf{F}$ and $\mathbf{g}$. Scalars are denoted by lowercase Roman and Greek letters, such as k and λ.

2. Literature review. To begin, let us define the problem under investigation:

DEFINITION 1. *Let* $\mathbf{M} \in \Re^{n\times n}$ *and* $\mathbf{q} \in \Re^n$. *The linear complementarity problem,* LCP($\mathbf{M}, \mathbf{q}$), *is to find an* $\mathbf{x} \in \Re^n$ *such that:*

$$\mathbf{x} \geq \mathbf{0}, \quad \mathbf{w} = \mathbf{Mx} + \mathbf{q} \geq \mathbf{0} \quad \textit{and} \quad \mathbf{w}^T\mathbf{x} = 0.$$

The following perturbed LCP (PLCP) will be used in the proposed algorithmic development.

DEFINITION 2. *Let* $\mathbf{M} \in \Re^{n\times n}$, $\mathbf{q} \in \Re^n$, *and* $\mu > 0$. *The perturbed* LCP, *denoted by* PLCP(μ), *is to find an* $\mathbf{x} \in \Re^n$ *such that:*

$$\mathbf{x} > \mathbf{0}, \quad \mathbf{w} = \mathbf{Mx} + \mathbf{q} > \mathbf{0} \quad \textit{and} \quad w_i x_i = \mu, \quad i = 1, \ldots, n.$$

In the study of this problem, the following definitions of various special matrices will be used.

DEFINITION 3. *The matrix* $\mathbf{M} \in \Re^{n\times n}$ *is said to be a*

1. *P_0-matrix if for all* $\mathbf{x} \neq \mathbf{0}$, *there exists an index i such that* $x_i \neq 0$ *and* $x_i(\mathbf{Mx})_i \geq 0$;
2. *positive semi-definite matrix if* $\mathbf{x}^T\mathbf{Mx} \geq 0$ *for all* $\mathbf{x}$;
3. *P-matrix if for all* $\mathbf{x} \neq 0$, *there exists an index i such that* $x_i(\mathbf{Mx})_i > 0$;
4. *positive definite matrix if* $\mathbf{x}^T\mathbf{Mx} > 0$ *for all* $\mathbf{x} \neq \mathbf{0}$;
5. *R_0-matrix if the following system has no nonzero solution:*

$$\begin{aligned}&\mathbf{x} \geq \mathbf{0},\\ &\mathbf{M}_{i\cdot}\mathbf{x} = 0 \ \textit{if}\ x_i > 0,\\ &\mathbf{M}_{i\cdot}\mathbf{x} \geq 0 \ \textit{if}\ x_i = 0;\end{aligned}$$

6. *Q-matrix if* LCP($\mathbf{M}, \mathbf{q}$) *has a solution for all* $\mathbf{q}$.

Among the above special matrices, the following relations hold:

positive definite matrix ⇒ P-matrix ⇒ P_0-matrix;

positive definite matrix ⇒ positive semidefinite matrix ⇒ P_0-matrix;

where ⇒ denotes implication.

The algorithms for the LCP fall into two broad classes: pivoting methods and interior-point techniques. The pivoting methods (Lemke, principal pivoting, etc.) are both well known and not directly related to the continuation method described in this paper; thus, we simply refer the reader to [21] for details. The only thing to note is that all these methods have an exponential worst-case performance in general.

The interior-point algorithms developed recently for the LCP are more akin to our method. Based on the idea of continuation or path following, Kojima et al. [14] first designed a polynomial algorithm for the monotone LCP, which computes a solution in $O(\sqrt{n}L)$ iterations, where n is the dimension of the LCP and L is the size of the input. The algorithm was then generalized by Kojima, Megiddo, and Noma [8] to solve nonlinear complementarity problems and LCPs with P_0-matrices. The potential reduction algorithms by Kojima, Mizuno, and Yoshise [13] and Ye [26] process an LCP with skew-symmetric or positive semidefinite matrices in $O(\sqrt{n}L)$ iterations; algorithms by Kojima, Megiddo, and Ye [9] and Ye [25] process an LCP with a P-matrix in $O(n^2 \max(|\lambda|/\gamma n, 1)L)$ iterations, where λ is the least eigenvalue of the symmetric part of $\mathbf{M}$ and γ is the P-matrix number defined by (see [18]):

$$\gamma = \arg\max_r \{\max_i x_i(\mathbf{Mx})_i \geq r\|\mathbf{x}\|^2 \quad \forall \mathbf{x} \in \Re^n\}.$$

The potential reduction algorithm has been further extended to a broader class of matrices by Ye [27]. The algorithm for monotone LCPs by Mizuno [19] includes both path following algorithms and potential reduction algorithms as its special cases. Finally, Kojima, Megiddo, Noma, and Yoshise [10] have recently carried out an extensive study of a unified approach for the path following and potential reduction algorithms.

Very little has been done in terms of numerical tests of the interior-point algorithms for the LCP. Mizuno, Yoshise, and Kikuchi [20] modified the path-following algorithm by Kojima et al. [14] for practical implementation and compared various path-following algorithms in a limited numerical experiment.

The path following algorithm for LCP$(\mathbf{M}, \mathbf{q})$ by Kojima et al. [14] is the method that is most closely related to the research presented in this paper. Assume $\mathbf{M}$ is positive semidefinite. The steps of the Kojima et al. algorithm are as follows:

Initiation Step. Construct an initial solution that satisfies:

1. $(\mathbf{x}^0)^T(\mathbf{w}^0) \leq 2^{O(L)}$
2. $(\mathbf{x}^0, \mathbf{w}^0)$ is close, as defined in [14], to the solution of PLCP(μ) for some μ.

Set $k = 1$ and go to Main Step.

Main Step

1. Let $\mu^k = \alpha^k \mu^{k-1}$, where $\alpha^k < 1$ is an appropriately defined parameter.
2. Compute the Newton direction $(\Delta\mathbf{x}^k, \Delta\mathbf{w}^k)$ at $(\mathbf{x}^k, \mathbf{w}^k)$ of the following systems of nonlinear equations:

$$\mathbf{w} = \mathbf{Mx} + \mathbf{q}, \quad w_i x_i = \mu^k, \quad i = 1, \ldots, n.$$

3. Let $\mathbf{x}^{k+1} = \mathbf{x}^k + \Delta\mathbf{x}^k$ and $\mathbf{w}^{k+1} = \mathbf{w}^k + \Delta\mathbf{w}^k$. If $(\mathbf{x}^{k+1})^T(\mathbf{w}^{k+1}) < 2^{-2L}$, stop; otherwise, set $k = k + 1$ and go to Step 1.

Kojima et al. proved that $(\mathbf{x}^{k+1})^T(\mathbf{w}^{k+1}) \leq \beta(\mathbf{x}^k)^T(\mathbf{w}^k)$ and $(\mathbf{x}^k, \mathbf{w}^k) > 0$ for all k and some $0 < \beta < 1$, if α^k is appropriately defined. Once $(\mathbf{x}^{k+1})^T(\mathbf{w}^{k+1}) < 2^{-2L}$, the solution of the LCP can be found in one step by solving a system of linear equations.

Other equation-based approaches for the LCP exist in the literature. Watson [24] presents a test of various homotopy methods, and Harker and Pang [7] study both theoretically and computationally the use of Newton's method for solving the LCP based on a nondifferentiable equation formulation of the problem.

3. The continuation method.

3. The continuation method. The algorithm proposed in this section is a direct application of the algorithm described in [3] for nonlinear complementarity problems (NCP). Thus, we refer to [3] for some of the proofs. We will show our algorithm has several advantages over path following interior-point algorithms. The algorithm is then extended to solve the general LCPs.

THEOREM 3.1. *Let $\mu > 0$ and $m_{ii} > 0$ for all i. Then $\mathbf{x}$ is a solution of* PLCP(μ) *if and only if it solves the following system of nonlinear equations, denoted by $\mathbf{J}(\mathbf{x}, \mu) = \mathbf{0}$:*

$$(1)\quad (\mathbf{Mx}+\mathbf{q})_i + m_{ii}x_i - \sqrt{[(\mathbf{Mx}+\mathbf{q})_i - m_{ii}x_i]^2 + 4m_{ii}\mu} = 0, \quad i = 1, \ldots, n.$$

Proof. See [3]. □

Denote

$$r_i(\mathbf{x}, \mu) = 1 - \frac{(\mathbf{Mx}+\mathbf{q})_i - m_{ii}x_i}{\sqrt{[(\mathbf{Mx}+\mathbf{q})_i - m_{ii}x_i]^2 + 4m_{ii}\mu}}.$$

Let $\mathbf{D} = Diag\{m_{ii}\}$ and $\mathbf{R}(\mathbf{x}, \mu) = Diag\{r_i(\mathbf{x}, \mu)\}$. Then, after some algebraic manipulation, we obtain

$$\nabla\mathbf{J}(\mathbf{x}, \mu) = \mathbf{R}(\mathbf{x}, \mu)[\mathbf{M} + (2\mathbf{R}^{-1}(\mathbf{x}, \mu) - \mathbf{I})\mathbf{D}].$$

Let $\mathbf{x}(\mu)$ be a solution of $\mathbf{J}(\mathbf{x}, \mu) = \mathbf{0}$ (see Theorem 3.7 for a sufficient condition for the existence of $x(\mu)$). Denote the set of paths by

$$(2)\qquad \mathbf{T} = \{(\mathbf{x}(\mu), \mu) : 0 < \mu \le \bar{\mu}\},$$

where $\bar{\mu}$ is some positive number.

THEOREM 3.2. *If $\nabla\mathbf{J}(\mathbf{x}(\mu), \mu)$ is nonsingular for all $\mathbf{x}(\mu)$, $0 < \mu \le \bar{\mu}$, then $\mathbf{T}$ consists solely of continuously differentiable paths. If, in addition, $\mathbf{T}$ is bounded, then $\mathbf{x}(\mu)$ approaches a limit point $\mathbf{x}(0) \in \mathbf{S}$ as μ approaches zero, where $\mathbf{S}$ is the solution set of the original* LCP.

Proof. The first part of the theorem is a direct application of the Path Theorem (see [28, p. 20]). The proof of the second part of the theorem is similar to that of [3, Thm. 5]. □

Therefore, if the conditions in Theorem 3.2 are satisfied, a path in the set $\mathbf{T}$ (which could contain multiple paths), if any exist, is continuously differentiable and leads to a solution of the original LCP. Various path following algorithms (see [28]) can then be used to trace the path to a solution of the LCP given that $\mathbf{x}(\mu)$ is known for some $0 < \mu \le \bar{\mu}$. In what follows, we first present a necessary and sufficient condition for $\nabla\mathbf{J}(\mathbf{x}, \mu)$ to be nonsingular for all $\mathbf{x}$ and $\mu > 0$. The result itself provides a new characterization of a P_0-matrix.

THEOREM 3.3. *$\mathbf{M}+\mathbf{D}$ is nonsingular for all positive diagonal matrices $\mathbf{D}$ if and only if $\mathbf{M}$ is a P_0-matrix.*

Lemmas 3.4 and 3.5 are established first in order to prove the above theorem.

LEMMA 3.4. *Let $\mathbf{X} = diag\{x_1, \ldots, x_n\}$, $x_i > 0$ for all $i = 1, \ldots, n$, be a positive diagonal matrix. If $\mathbf{M} + \mathbf{X}$ is nonsingular for any $\mathbf{X}$, then $|\mathbf{M} + \mathbf{X}| > 0$ for all $\mathbf{X}$.*

Proof. Denote $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $f(\mathbf{x}) = |\mathbf{M}+\mathbf{X}|$. Then $f(\mathbf{x})$ is a polynomial function of $\mathbf{x}$. We first show that either $f(\mathbf{x}) < 0$ or $f(\mathbf{x}) > 0$. Suppose, on the contrary, that there are two positive diagonal matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ such that $f(\mathbf{x}_1) < 0$

and $f(\mathbf{x}_2) > 0$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the corresponding vectors of matrices $\mathbf{X}_1$ and $\mathbf{X}_2$. Then by the mean value theorem, there exists a positive diagonal matrix $\mathbf{X}$ such that $f(\mathbf{x}) = |\mathbf{M}+\mathbf{X}| = 0$. However, this violates the assumption of the lemma that $\mathbf{M}+\mathbf{X}$ is nonsingular. We now show that $f(\mathbf{x}) > 0$. It is not difficult to see that as all x_i approaches $+\infty$, the expansion term $\prod_{i=1}^{n} x_i$ of $f(\mathbf{x})$ becomes dominant. Therefore, there exists a $\delta > 0$ such that $f(\mathbf{x}) = |\mathbf{M}+\mathbf{X}| > 0$ if $x_i \geq \delta$ for all $i = 1, \ldots, n$. □

LEMMA 3.5. *Let* $\mathbf{X}$ *be defined as in Lemma* 3.4 *and* $\mathbf{Y} = diag\{y_1, \ldots, y_n\}$, $y_i \geq 0$ *for all* $i = 1, \ldots, n$, *be a nonnegative, diagonal matrix. If* $\mathbf{M} + \mathbf{X}$ *is nonsingular for any* $\mathbf{X}$, *then* $|\mathbf{M} + \mathbf{Y}| \geq 0$ *for all* $\mathbf{Y}$.

Proof. Denote $I = \{i : y_i = 0\}$ and $I \cup \bar{I} = \{1, \ldots, n\}$. Let $x_i = y_i$ for all $i \in \bar{I}$ and $x_i > 0$ for all $i \in I$. Then

$$|\mathbf{M}+\mathbf{Y}| = \lim_{x_{i\in I} \to 0^+} |\mathbf{M}+\mathbf{X}| \geq 0,$$

where the equality follows from the continuity of $|\mathbf{M} + \mathbf{X}|$ and the inequality follows from Lemma 3.4. □

Proof of Theorem 3.3. The sufficiency part is well known, and $|\mathbf{M}| \geq 0$ by Lemma 3.5. Now let I be any subset of $\{1, \ldots, n\}$ and $\bar{I}$ be its complement. We show that $|\mathbf{M}_{II}| \geq 0$, where $\mathbf{M}_{II}$ is the submatrix of $\mathbf{M}$ defined by the index set I. Suppose on the contrary that $|\mathbf{M}_{II}| < 0$. Let $\mathbf{X}$ be a diagonal matrix such that

$$x_i = 0 \quad \forall i \in I \quad \text{and} \quad x_i > 0 \quad \forall i \in \bar{I}.$$

The expansion term $|\mathbf{M}_{II}| \prod_{i\in\bar{I}} x_i$ of $f(\mathbf{x})$ becomes dominant as $x_{i\in\bar{I}}$ approaches $+\infty$. Therefore, there exists a $\delta > 0$ such that $|\mathbf{M} + \mathbf{X}| < 0$ when $x_i \geq \delta$ for all $i \in \bar{I}$. However, this contradicts the fact that $|\mathbf{M} + \mathbf{X}| \geq 0$. Therefore, $|\mathbf{M}_{II}| \geq 0$ for all $I \subset \{1, \ldots, n\}$ and thus, $\mathbf{M}$ is a P_0-matrix [6]. □

The following proposition is a direct consequence of Theorem 3.3.

PROPOSITION 3.6. *Assume* $m_{ii} > 0$ *for all* $i = 1, \ldots, n$. $\nabla\mathbf{J}(\mathbf{x}, \mu)$ *is nonsingular for all* $\mathbf{x} \in \Re^n$, *and* $\mu > 0$ *if* $\mathbf{M}$ *is a* P_0*-matrix. Assume, in addition, that* $\mathbf{M} - \mathbf{D}$ *is nonsingular, then the conclusion is true only if* $\mathbf{M}$ *is a* P_0*-matrix.*

Proof. One can verify that $0 < r_i(\mathbf{x}, \mu) < 2$ for all $\mu > 0$ and $i = 1, \ldots, n$. As a result, both $\mathbf{R}(\mathbf{x}, \mu)$ and $(2\mathbf{R}^{-1}(\mathbf{x}, \mu) - \mathbf{I})\mathbf{D}$ are positive diagonal matrices since $\mathbf{D} > 0$ by assumption. Therefore, the second part of the Jacobian, and thus the Jacobian itself, is nonsingular if $\mathbf{M}$ is a P_0-matrix. To establish necessity, suppose that $\mathbf{M}$ is not a P_0-matrix. By Theorem 3.3, there exists a positive diagonal matrix $\mathbf{L} = diag\{l_1, \ldots, l_n\}$ such that $\mathbf{M} + \mathbf{L}$ is singular. Choose $\mathbf{x} = (\mathbf{M} - \mathbf{D})^{-1}\mathbf{t}$, where $\mathbf{t}$ is a vector whose elements are given by

$$t_i = (l_i - m_{ii})\sqrt{\frac{\mu}{l_i}} - q_i \quad \forall i = 1, \ldots, n.$$

One can verify that $(2\mathbf{R}^{-1}(\mathbf{x}, \mu) - \mathbf{I})\mathbf{D} = \mathbf{L}$ and that the Jacobian $\nabla\mathbf{J}(\mathbf{x}, \mu)$ is singular. □

Although any LCP defined by a matrix with a positive diagonal can be formulated as a system of nonlinear equations $\mathbf{J}(\mathbf{x}, \mu) = \mathbf{0}$ and then solved by a path following algorithm, Proposition 3.6 states that the most general LCP that can be solved without trouble is the LCP with a P_0-matrix. When $\mathbf{M}$ is positive semidefinite, it has been shown [14] that PLCP(μ) has a unique solution for all $\mu > 0$ given that the original LCP has a strictly feasible solution. When $\mathbf{M}$ is a P_0-matrix, PLCP(μ) does not necessarily have a solution even if the LCP has a solution. An example is the LCP defined by

$$\mathbf{M} = \begin{pmatrix} 2 & -1 \\ -4 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{q} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

If the LCP does not have a solution, then PLCP(μ) may have no solution for all $\mu > 0$, or may have a solution for some $\mu > \bar{\mu}$ and no solution for $0 < \mu \leq \bar{\mu}$. An example would be the LCP defined by

$$\mathbf{M} = \begin{pmatrix} 0 & 1 \\ 0 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{q} = \begin{pmatrix} -1 \\ -1 \end{pmatrix},$$

where the PLCP(μ) has solutions for all $\mu > 1$ and no solution for $0 < \mu \leq 1$. In cases where PLCP(μ) has solutions for all $0 < \mu$ (we did not find such examples), the solution $\mathbf{x}(\mu)$ must be unbounded as μ approaches zero. If this is not true, by Theorem 3.2 (see also [12, Thm. 3.1]), $\mathbf{x}(\mu)$ will approach a solution of the LCP, which is a contradiction. The following result provides a sufficient condition for PLCP(μ) to have a solution. We remark that different conditions have been proposed by Kojima et al. [8] for this result to hold.

THEOREM 3.7. *Let* $\mathbf{M} \in P_0$ *and* $m_{ii} > 0$ *for all* i. *If there exists an* $\mathbf{x}^0$ *such that the set*

$$X(\mu) = \{\mathbf{x} : \|\mathbf{J}(\mathbf{x}, \mu)\| \leq \|\mathbf{J}(\mathbf{x}^0, \mu)\|\}$$

is uniformly bounded for all $0 < \mu \leq \bar{\mu}$ *and some* $\bar{\mu} > 0$, *then* PLCP(μ) *has a unique solution* $\mathbf{x}(\mu)$ *for all* $0 < \mu \leq \bar{\mu}$ *and* $\mathbf{x}(\mu)$ *approaches a solution of the* LCP *as* μ *approaches zero.*

Proof. Let $\mathbf{x}^k$ be a sequence generated by damped Newton method[1] for equation $\mathbf{J}(\mathbf{x}, \mu) = \mathbf{0}$, starting from $\mathbf{x}^0 \in X(\mu)$ for some $0 < \mu \leq \bar{\mu}$. In general, let $\mathbf{x}^k \in X(\mu)$. $\nabla \mathbf{J}(\mathbf{x}^k, \mu)$ is nonsingular by Proposition 3.6 since $\mathbf{M} \in P_0$. Therefore, $\mathbf{x}^{k+1}$ is well defined and, in addition, $\mathbf{x}^{k+1} \in X(\mu)$ because

$$\|\mathbf{J}(\mathbf{x}^{k+1}, \mu)\| \leq \|\mathbf{J}(\mathbf{x}^k, \mu)\| \leq \|\mathbf{J}(\mathbf{x}^0, \mu)\|.$$

Since $\mathbf{x}^0 \in X(\mu)$, by induction, the entire sequence $\{\mathbf{x}^k\}$ generated by damped Newton's method is well defined and bounded. Therefore, $\{\mathbf{x}^k\}$ has an accumulation point $\mathbf{x}^* \in X(\mu)$. Applying the Global Convergence Theorem by Zangwill (see [16, p. 187]), one can show that $\mathbf{J}(\mathbf{x}^*, \mu) = \mathbf{0}$. Therefore, $\mathbf{x}^* = \mathbf{x}(\mu)$, the solution of $\mathbf{J}(\mathbf{x}, \mu) = \mathbf{0}$. The uniqueness of the solution is established in [3, Thm. 15] for the perturbed nonlinear complementarity problem, PNCP(μ). Since the set $X(\mu)$ is uniformly bounded for all $0 < \mu \leq \bar{\mu}$ by assumption, $\mathbf{x}(\mu) = \mathbf{x}^* \in X(\mu)$ is uniformly bounded for all $0 < \mu \leq \bar{\mu}$ and the path $\mathbf{T}$ defined by (2) is bounded. By Theorem 3.2, $\mathbf{x}(\mu)$ approaches a solution of the LCP as μ approaches zero. □

Suppose an LCP satisfies the conditions of Theorem 3.7, then a solution of the LCP can be obtained by a path following algorithm. In addition, from the proof of Theorem 3.7, it is clear that the solution of PLCP(μ) can be found by starting from any initial point. Therefore, the path does not have to be followed very closely when μ is relatively large. We now show that the conditions of Theorem 3.7 are satisfied if

[1] Damped Newton method is a globalized Newton method by incorporating line search using $\|\mathbf{J}\|^2$ as the merit function. See Algorithm 1 for detailed description.

$\mathbf{M} \in P_0 \cap R_0$. It was shown in [1] that $\mathbf{M} \in P_0 \cap R_0$ implies $\mathbf{M} \in Q$; i.e., the LCP has a solution for all $\mathbf{q} \in \Re^n$.

LEMMA 3.8. *If* $\mathbf{M} \in R_0$ *and* $m_{ii} > 0$ *for all* i, *then the set*

$$X(\mu, \alpha) = \{\mathbf{x} : \|\mathbf{J}(\mathbf{x}, \mu)\| \le \alpha\}$$

is uniformly bounded for all $0 \le \mu \le \bar{\mu} < +\infty$, *and* $0 \le \alpha \le \bar{\alpha} < +\infty$.

Proof. Define $F_i(\mathbf{x}_{-i}) = \sum_{j \ne i} m_{ij} x_j + q_i$ and $\mathbf{F}(\mathbf{x}) = (F_1(\mathbf{x}_{-1}), \ldots, F_n(\mathbf{x}_{-n}))$. Then

$$\begin{aligned} J_i(\mathbf{x}, \mu) &= 2m_{ii}x_i + F_i(\mathbf{x}_{-i}) - \sqrt{F_i^2(\mathbf{x}_{-i}) + 4\mu m_{ii}} \\ &= 2\mathbf{M}_{i\cdot}\mathbf{x} + 2q_i - F_i(\mathbf{x}_{-i}) - \sqrt{F_i^2(\mathbf{x}_{-i}) + 4\mu m_{ii}}. \end{aligned} \tag{3}$$

Suppose the set $X(\mu, \alpha)$ is not uniformly bounded. Then there exists an unbounded sequence $\{\mathbf{x}^k, k \in \mathcal{N}\}$ such that

$$\mathbf{x}^k \in X(\mu^k, \alpha^k), \quad \|\mathbf{x}^k\| \to +\infty, \quad \mu^k \to \tilde{\mu}, \quad \text{and } \alpha^k \to \tilde{\alpha},$$

where $0 \le \mu^k, \tilde{\mu} \le \bar{\mu}$, and $0 \le \alpha^k, \tilde{\alpha} \le \bar{\alpha}$. One can choose an unbounded subsequence $\mathcal{K} \subseteq \mathcal{N}$ such that for each $i \in \{1, \ldots, n\}$ the sequence $\{F_i(\mathbf{x}_{-i}^k),\ k \in \mathcal{K}\}$ is either bounded, or unbounded from above, or unbounded from below. Let

$$\begin{aligned} S_+ &= \{i : \lim_{k \to \infty} F_i(\mathbf{x}_{-i}^k) = +\infty, \quad k \in \mathcal{K}\}, \\ S_- &= \{i : \lim_{k \to \infty} F_i(\mathbf{x}_{-i}^k) = -\infty, \quad k \in \mathcal{K}\}, \\ S_0 &= \{i : -\infty < \lim_{k \to \infty} F_i(\mathbf{x}_{-i}^k) < +\infty, \quad k \in \mathcal{K}\}. \end{aligned}$$

To lead to a contradiction, we show that either the subsequence $\{\mathbf{x}^k,\ k \in \mathcal{K}\}$ is bounded or $\mathbf{M}$ cannot be a R_0 matrix. We assume $k \in \mathcal{K}$ in the rest of the proof.

If $i \in S_+ \cup S_0$, there exists a $-\infty < B \le 0$ such that

$$B \le F_i(\mathbf{x}_{-i}^k) - \sqrt{F_i^2(\mathbf{x}_{-i}^k) + 4\mu^k m_{ii}} \le 0$$

for all $0 \le \mu^k < \bar{\mu}$ since $m_{ii} > 0$. This implies that x_i^k is bounded since $m_{ii} > 0$ and $|J_i(\mathbf{x}^k, \mu^k)| \le \|\mathbf{J}(\mathbf{x}^k, \mu^k)\| \le \alpha^k \le \bar{\alpha}$. Hence, if $S_- = \emptyset$, $\mathbf{x}^k$ is a bounded sequence, which contradicts the assumption.

Let $S_- \ne \emptyset$ and $i \in S_-$, then

$$F_i(\mathbf{x}_{-i}^k) - \sqrt{F_i^2(\mathbf{x}_{-i}^k) + 4\mu^k m_{ii}} \to -\infty.$$

Therefore, $x_i^k \to +\infty$ since $m_{ii} > 0$ and $|J_i(\mathbf{x}^k, \mu^k)| \le \bar{\alpha}$. Define a new sequence

$$\mathbf{y}^k = \frac{\mathbf{x}^k}{\|\mathbf{x}^k\|},$$

and let $\tilde{\mathbf{y}}$ be its limit point. From the above discussion, x_i^k is bounded for all $i \in S_+ \cup S_0$ and approaches $+\infty$ when $i \in S_-$. Therefore,

$$\tilde{y}_i \begin{cases} \ge 0 & \text{if } i \in S_- \\ = 0 & \text{if } i \in S_+ \cup S_0. \end{cases}$$

Equation (3) can be rewritten as

$$(4)\quad \frac{1}{\|\mathbf{x}^k\|}J_i(\mathbf{x}^k,\mu^k) = 2\mathbf{M}_{i\cdot}\mathbf{y}^k + \frac{2q_i}{\|\mathbf{x}^k\|} - \frac{1}{\|\mathbf{x}^k\|}\left[F_i(\mathbf{x}^k_{-i}) + \sqrt{F_i^2(\mathbf{x}^k_{-i}) + 4\mu^k m_{ii}}\right].$$

Passing the limit on both sides as $k \to \infty$, the left-hand side becomes zero since $|J_i(\mathbf{x}^k,\mu^k)| \leq \bar{\alpha}$. Notice that $F_i(\mathbf{x}^k_{-i}) + \sqrt{F_i^2(\mathbf{x}^k_{-i}) + 4\mu^k m_{ii}}$ is bounded for all $0 \leq \mu^k < \bar{\mu}$ if $F_i(\mathbf{x}^k_{-i})$ is negative or bounded, and it approaches $+\infty$ if $F_i(\mathbf{x}^k_{-i})$ approaches $+\infty$. Therefore,

$$\lim_{k\to\infty}\frac{1}{\|\mathbf{x}^k\|}\left[F_i(\mathbf{x}^k_{-i}) + \sqrt{F_i^2(\mathbf{x}^k_{-i}) + 4\mu^k m_{ii}}\right]\begin{cases} \geq 0 & \text{if } i \in S_+ \\ = 0 & \text{if } i \in S_- \cup S_0. \end{cases}$$

Thus, identity (4) implies

$$\mathbf{M}_{i\cdot}\tilde{y}\begin{cases} \geq 0 & \text{if } i \in S_+ \\ = 0 & \text{if } i \in S_- \cup S_0. \end{cases}$$

Define

$$S = \{i \in S_- : \tilde{y}_i > 0\}.$$

Then, $\tilde{\mathbf{y}} \geq 0$ satisfies

$$\begin{aligned} &\tilde{y}_i > 0, \mathbf{M}_{i\cdot}\tilde{\mathbf{y}} = 0 \quad \forall i \in S \subset S_-, \\ &\tilde{y}_i = 0, \mathbf{M}_{i\cdot}\tilde{\mathbf{y}} \geq 0 \quad \forall i \in \{1,\dots,n\}\backslash S. \end{aligned}$$

However, this contradicts the assumption that $\mathbf{M} \in R_0$. □

Combining Lemma 3.8 and Theorem 3.7, we obtain the following existence result for PLCP(μ).

COROLLARY 3.9. *Suppose* $\mathbf{M} \in P_0 \cap R_0$ *and* $m_{ii} > 0$ *for all* i. *Then* PLCP(μ) *has a unique solution for all* $\mu > 0$ *and the solution approaches one of the solutions of the* LCP *as* μ *approaches zero.*

Corollary 3.9 also provides an alternative proof for the existence of a solution for the LCP with $\mathbf{M} \in P_0 \cap R_0$. We now present our continuation method for the LCP with a matrix $\mathbf{M} \in P_0 \cap R_0$ and positive diagonal elements.

Initiation Step. Let ε be a given stopping tolerance. Choose an initial point $\mathbf{x}^0 \in \Re^n$ and sequences $\mu^k > 0$ and $\epsilon^k > 0$, $k = 0, 1, 2, \dots$ such that

$$\begin{aligned} \mu^k < \mu^{k-1} \quad &\text{and} \quad \lim_{k\to\infty}\mu^k = 0, \\ \epsilon^k < \epsilon^{k-1} \quad &\text{and} \quad \lim_{k\to\infty}\epsilon^k = 0. \end{aligned}$$

Set $k = 0$ and go to the Main Step.

Main Step

1. Starting with $\mathbf{x}^{k-1}$, find $\mathbf{x}^k$ by using Newton's method to solve $\mathbf{J}(\mathbf{x},\mu^k) = \mathbf{0}$ up to the point where $\|\mathbf{J}(\mathbf{x}^k,\mu^k)\| \leq \epsilon^k$.
2. If $\|\min\{\mathbf{x}^k,\mathbf{w}^k\}\| \leq \varepsilon$, where $\mathbf{w} = \mathbf{M}\mathbf{x} + \mathbf{q}$ and "min" is taken componentwise, terminate; otherwise, go to Step 1.

In what follows, we will compare our algorithm with the Polynomial Path Following (PPF) algorithm proposed by Kojima et al. [14]. The proposed continuation method requires $m_{ii} > 0$ for all $i = 1, \ldots, n$, which is not necessary in the PPF. However, this assumption is satisfied or can be avoided in many cases such as the symmetric monotone LCP, QP, and LP as discussed in [4]. Both the PPF and the continuation method follow the same path; the difference is in the representation of the path. In the PPF, the path is represented by PLCP(μ), and a new point is obtained by applying Newton's method to the following system of nonlinear equations:

$$(\mathbf{Mx} + \mathbf{q})_i x_i = \mu, \qquad i = 1, \ldots, n,$$

which, by factorization, can be rewritten as

$$\left[(\mathbf{Mx} + \mathbf{q})_i + m_{ii}x_i + \sqrt{[(\mathbf{Mx} + \mathbf{q})_i - m_{ii}x_i]^2 + 4m_{ii}\mu}\right] J_i(\mathbf{x}, \mu) = 0, \quad i = 1, \ldots, n,$$

where $J_i(\mathbf{x}, \mu)$ is defined in (1). One can verify that at the solution of PLCP(μ), the first part of the right-hand side of the equation is positive. By dropping this part, we obtain a new representation of the path $\mathbf{J}(\mathbf{x}, \mu) = \mathbf{0}$, which is used in the proposed continuation method. Geometrically, the Newton direction obtained by the PPF is a compromise among the Newton directions of all 2^n paths that lead to the complementarity solutions of

$$(\mathbf{Mx} + \mathbf{q})_i x_i = 0, \quad i = 1, \ldots, n.$$

Among these 2^n paths, only one satisfies the feasibility condition and our algorithm always follows this very path. As a consequence of the above comparison, we believe that our algorithm will outperform the PPF computationally. The time saving will come from the following factors:

1. The continuation method identifies the right path to follow and we believe the Newton direction is more "accurate." More research has to be done in order to claim this point rigorously.
2. Our algorithm can start from any point including those in the negative orthant, while the PPF and other interior-point algorithms have to construct an initial point which is usually far away from the solution of the problem. For certain problems, such as the LCP with a row sufficient matrix or P_0-matrix, the construction of the initial solution dramatically increases the dimension of the problem and reduces the efficiency of the path following algorithm.
3. Since our algorithm does not require that each intermediate iterate stay feasible, it is possible for our algorithm to take a larger step size at each iteration.

Numerically, both algorithms are ill conditioned as μ approaches zero. Let $\mathbf{w} = \mathbf{Mx} + \mathbf{q}$ and $\mathbf{X}$, $\mathbf{W}$ be two diagonal matrices whose ith entries are x_i and w_i, respectively. In the PPF, the matrix to be inverted is $\mathbf{M} + \mathbf{WX}^{-1}$, while in our algorithm, it is $\mathbf{M} + (2\mathbf{R}^{-1}(\mathbf{x}, \mu) - \mathbf{I})\mathbf{D}$. Therefore, both algorithms invert matrices of the same structure at each iteration. In addition, the following result shows that the above two matrices are identical at the solution of PLCP(μ), or $\mathbf{J}(\mathbf{x}, \mu) = \mathbf{0}$.

PROPOSITION 3.10. *Suppose* $\mathbf{x}^*$ *is a solution of* PLCP(μ), *then*

$$\mathbf{M} + \mathbf{W}^*\mathbf{X}^{*-1} = \mathbf{M} + (2\mathbf{R}^{-1}(\mathbf{x}^*, \mu) - \mathbf{I})\mathbf{D},$$

where $\mathbf{W}^*$ *and* $\mathbf{X}^*$ *are the corresponding diagonal matrices defined at* $(\mathbf{x}^*, \mathbf{w}^*)$.

Proof. It suffices to show that $w_i^*/x_i^* = (2/r_i(\mathbf{x}^*, \mu) - 1)m_{ii}$:

$$\begin{aligned} r_i(\mathbf{x}^*, \mu) &= 1 - \frac{w_i^* - m_{ii}x_i^*}{\sqrt{(w_i^* - m_{ii}x_i^*)^2 + 4m_{ii}\mu}} \\ &= 1 - \frac{\mu - m_{ii}(x_i^*)^2}{\sqrt{(\mu - m_{ii}(x_i^*)^2)^2 + 4m_{ii}\mu(x_i^*)^2}} \\ &= 1 - \frac{\mu - m_{ii}(x_i^*)^2}{\mu + m_{ii}(x_i^*)^2} \\ &= \frac{2m_{ii}(x_i^*)^2}{\mu + m_{ii}(x_i^*)^2}. \end{aligned}$$

Therefore,

$$\left(\frac{2}{r_i(\mathbf{x}^*, \mu)} - 1\right) m_{ii} = \frac{\mu}{(x_i^*)^2} = \frac{w_i^*}{x_i^*}. \qquad \square$$

This result roughly demonstrates that the PPF algorithm and our algorithm have the same degree of ill-conditioning as μ goes to zero.

Now we will extend the above ideas to the general LCP. Consider a matrix $\mathbf{N} \in \Re^{n \times n}$ that is assumed to have no special structure and whose diagonal could have negative elements. The general LCP, denoted by LCP$(\mathbf{N}, \mathbf{q})$, is to find an $\mathbf{x}$ such that

$$\mathbf{z} = \mathbf{N}\mathbf{x} + \mathbf{q} \geq \mathbf{0}, \quad \mathbf{x} \geq \mathbf{0} \quad \text{and} \quad \mathbf{x}^T\mathbf{z} = 0.$$

The above problem is NP-complete in general. Choose a λ such that $\mathbf{M} = \mathbf{N} + \lambda\mathbf{I}$ is positive definite. We formulate the following problem, denoted by LCP$(\mathbf{M}, \mathbf{q}, \lambda)$, of finding an $\mathbf{x}$ such that

$$\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{q} \geq \mathbf{0}, \quad \mathbf{x} \geq \mathbf{0} \quad \text{and} \quad x_i y_i = \lambda x_i^2, \quad i = 1, \ldots, n.$$

PROPOSITION 3.11. $(\mathbf{x}^*, \mathbf{y}^*)$ *solves* LCP$(\mathbf{M}, \mathbf{q}, \lambda)$ *if and only if* $(\mathbf{x}^*, \mathbf{z}^*)$ *solves* LCP$(\mathbf{N}, \mathbf{q})$, *where* $\mathbf{z}^* = \mathbf{y}^* - \lambda\mathbf{x}^*$.

Proof. It is clear that $\mathbf{z}^* = \mathbf{N}\mathbf{x}^* + \mathbf{q}$ if and only if $\mathbf{y}^* = \mathbf{M}\mathbf{x}^* + \mathbf{q}$ by the construction of $\mathbf{M}$ and $\mathbf{z}^*$. Suppose $(\mathbf{x}^*, \mathbf{z}^*)$ solves LCP$(\mathbf{N}, \mathbf{q})$. Then

$$\begin{aligned} \mathbf{y}^* &= \mathbf{z}^* + \lambda\mathbf{x}^* \geq \mathbf{0}, \\ x_i^* y_i^* &= x_i^* z_i^* + \lambda x_i^{*2} = \lambda x_i^{*2}. \end{aligned}$$

Hence, $(\mathbf{x}^*, \mathbf{y}^*)$ solves LCP$(\mathbf{M}, \mathbf{q}, \lambda)$. Suppose now that $(\mathbf{x}^*, \mathbf{y}^*)$ solves LCP$(\mathbf{M}, \mathbf{q}, \lambda)$. That $\mathbf{z}^* \geq \mathbf{0}$ is shown case by case:

1. If $y_i^* = 0$, then $x_i^* = 0$ by the equality $x_i^* y_i^* = \lambda x_i^{*2}$. Therefore, $\mathbf{z}_i^* = \mathbf{y}_i^* - \lambda\mathbf{x}_i^* = 0$.
2. If $y_i^* > 0$ and $x_i^* = 0$, then $z_i^* = y_i^* > 0$.
3. If $y_i^* > 0$ and $x_i^* > 0$, then $y_i^* = \lambda x_i^*$ from the equality $x_i^* y_i^* = \lambda x_i^{*2}$. Therefore, $z_i^* = y_i^* - \lambda x_i^* = 0$.

Therefore, $(\mathbf{x}^*, \mathbf{z}^*)$ solves LCP$(\mathbf{N}, \mathbf{q})$. $\square$

Associated with LCP$(\mathbf{M}, \mathbf{q}, \lambda)$, we define the following perturbed problem, denoted by PLCP$(\mathbf{M}, \mathbf{q}, \lambda, \mu)$, of finding an $\mathbf{x}$ such that

$$\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{q} > \mathbf{0}, \quad \mathbf{x} > \mathbf{0} \quad \text{and} \quad x_i y_i = \lambda x_i^2 + \mu, \quad i = 1, \ldots, n.$$

Similar to Theorem 3.1, we have the following proposition.

PROPOSITION 3.12. *The vector* $\mathbf{x}$ *is a solution of* PLCP$(\mathbf{M}, \mathbf{q}, \lambda, \mu)$ *if and only if it solves the following system of nonlinear equations, denoted by* $\mathbf{J}(\mathbf{x}, \lambda, \mu) = \mathbf{0}$*:*

$$(\mathbf{Mx}+\mathbf{q})_i + m_{ii}x_i - \sqrt{[(\mathbf{Mx}+\mathbf{q})_i - m_{ii}x_i]^2 + 4m_{ii}(\lambda x_i^2 + \mu)} = 0,$$

or equivalently,

$$\begin{aligned}&((\mathbf{N}+\lambda\mathbf{I})\mathbf{x}+\mathbf{q})_i + (n_{ii}+\lambda)x_i \\ &\quad - \sqrt{[((\mathbf{N}+\lambda\mathbf{I})\mathbf{x}+\mathbf{q})_i - (n_{ii}+\lambda)x_i]^2 + 4(n_{ii}+\lambda)(\lambda x_i^2+\mu)} = 0,\end{aligned}$$

where $\mathbf{M} = \mathbf{N} + \lambda\mathbf{I}$.

Denote

$$r_i(\mathbf{x}, \lambda, \mu) = 1 - \frac{((\mathbf{N}+\lambda\mathbf{I})\mathbf{x}+\mathbf{q})_i - (n_{ii}+\lambda)x_i}{\sqrt{[((\mathbf{N}+\lambda\mathbf{I})\mathbf{x}+\mathbf{q})_i - (n_{ii}+\lambda)x_i]^2 + 4(n_{ii}+\lambda)(\lambda x_i^2+\mu)}}$$

$$t_i(\mathbf{x}, \lambda, \mu) = 1 - \frac{2\lambda x_i}{\sqrt{[((\mathbf{N}+\lambda\mathbf{I})\mathbf{x}+\mathbf{q})_i - (n_{ii}+\lambda)x_i]^2 + 4(n_{ii}+\lambda)(\lambda x_i^2+\mu)}}.$$

Let $\mathbf{D}_N$, $\mathbf{R}(\mathbf{x}, \lambda, \mu)$, and $\mathbf{T}(\mathbf{x}, \lambda, \mu)$ be diagonal matrices whose entries are given by n_{ii}, $r_i(\mathbf{x}, \lambda, \mu)$, and $t_i(\mathbf{x}, \lambda, \mu)$, respectively. After some algebraic manipulation, we obtain

$$\nabla\mathbf{J}(\mathbf{x}, \lambda, \mu) = \mathbf{R}(\mathbf{x}, \lambda, \mu)\{\mathbf{N} + \lambda\mathbf{I} + (\mathbf{D}_N + \lambda\mathbf{I})[2\mathbf{T}(\mathbf{x}, \lambda, \mu)\mathbf{R}^{-1}(\mathbf{x}, \lambda, \mu) - \mathbf{I})]\}.$$

In general, $\nabla\mathbf{J}(\mathbf{x}, \lambda, \mu)$ may be singular, even at the solution point of $\mathbf{J}(\mathbf{x}, \lambda, \mu) = \mathbf{0}$. Therefore, the performance of the continuation algorithm cannot be guaranteed. However, the algorithm provides one way to solve this difficult problem in an equation setting.

4. Implementation of the continuation method.

We have shown in §3 that PLCP(μ) and $\mathbf{J}(\mathbf{x}, \mu)$ share the same solutions if the matrix $\mathbf{M}$ has a positive diagonal. In addition, if $\mathbf{M} \in P_0 \cap R_0$ and $m_{ii} > 0$, the solution of $\mathbf{J}(\mathbf{x}, \mu)$ approaches that of the original LCP as μ goes to zero. Therefore, the solution of an LCP can be approximated by that of $\mathbf{J}(\mathbf{x}, \mu)$ with a very small μ; the smaller the μ, the better the approximation. However, as μ goes to zero, the Jacobian of $\mathbf{J}(\mathbf{x}, \mu)$ becomes increasingly ill conditioned. To reduce the ill conditioning, the method starts with a fairly large μ and decreases it while solving the nonlinear system of equations.

We now present the implementation details of the continuation method for the LCP with a positive semidefinite matrix or P_0-matrix. Define the error associated with the LCP at a point $\mathbf{x}^k$ as

$$err(\mathbf{x}^k) = \|\min\{\mathbf{x}^k, \mathbf{w}^k\}\|,$$

where $\mathbf{w}^k = \mathbf{Mx}^k + \mathbf{q}$ and "min" is taken componentwise. The detailed implementation of the continuation method is stated as follows:

ALGORITHM 1 (CONTINUATION METHOD FOR LCPS). *Let* $\mathbf{x}^0$ *be an arbitrary initial vector and* $\mu^0 > 0$. *Let* s, β, *and* σ *be given scalars with* $s > 0$, $\beta \in (0, 1)$, *and* $\sigma \in (0, 1/2)$. *In general, given* $\mathbf{x}^k$ *with* $\mathbf{J}(\mathbf{x}^k, \mu^k) \neq 0$, *we obtain the next iterate* $\mathbf{x}^{k+1}$ *by the following steps:*

Step 1. *Solve the Newton equation*

$$\mathbf{J}(\mathbf{x}^k, \mu^k) + \nabla\mathbf{J}(\mathbf{x}^k, \mu^k)\mathbf{d}^k = \mathbf{0}$$

for the direction $\mathbf{d}^k$.

Step 2. *Let* $\lambda_k = \beta^{m_k}s$, *where* m_k *is the smallest integer* m *for which the following condition holds:*

$$(1 - 2\sigma\beta^m s)\|\mathbf{J}(\mathbf{x}^k, \mu^k)\|^2 \geq \|\mathbf{J}(\mathbf{x}^k + \beta^m s\mathbf{d}^k, \mu^k)\|^2. \tag{5}$$

Set $\mathbf{x}^{k+1} = \mathbf{x}^k + \lambda_k \mathbf{d}^k$ *and check if* $err(\mathbf{x}^{k+1}) \leq \epsilon$, *where* ϵ *is a small positive constant.*

Step 3. *If the convergence test is not satisfied, find* μ^{k+1} *based on* μ^k, $err(\mathbf{x}^{k+1})$ *and* $\|\mathbf{J}(\mathbf{x}^{k+1}, \mu^k)\|$.

By Theorem 3.7, $\mathbf{J}(\mathbf{x}, \mu)$ is nonsingular for all $\mathbf{x} \in X(\mu)$ and thus, $\mathbf{d}^k \neq \mathbf{0}$ for all k. This together with the fact that $\|\mathbf{J}(\mathbf{x}, \mu)\|$ is bounded below guarantees the existence of m in step (5) (see [2, p. 22]). Algorithm 1 is now shown to have properties similar to interior-point algorithms for an LP when $\mathbf{M}$ is a positive semidefinite or P_0-matrix. Specifically, the Jacobian of $\mathbf{J}(\mathbf{x}, \mu)$ for all $\mathbf{x}$ and μ is of the following form:

$$\nabla\mathbf{J}(\mathbf{x}, \mu) = \mathbf{R}(\mathbf{x}, \mu)[\mathbf{M} + (2\mathbf{R}^{-1}(\mathbf{x}, \mu) - \mathbf{I})\mathbf{D}],$$

and the second part in the bracket is a positive diagonal matrix. Therefore, the matrix to be inverted at each iteration has the same nonzero structure. This property allows us to apply minimum reordering heuristics and symbolic factorization techniques to reduce fill-in during the numerical factorization of the matrix. Although these techniques apply to LCPs with general positive semidefinite and P_0-matrices, we concentrate our tests on problems in which $\mathbf{M}$ is symmetric positive semidefinite, since in this case the minimum reordering heuristic is more efficient and Cholesky factorization can be used to solve the linear equations [5]. In our implementation, the multiple minimum ordering heuristic [15] is performed first, and the linear equations are then solved numerically by SPARSPAK [5]. If the matrix $\mathbf{M}$ is 100% dense, the linear equations are solved directly by routines DGEFA-DGESL from LINPACK. In both cases, the Armijio line search procedure described in Step 2 is employed.

As in the case of the path following algorithm for the LP [17], it is crucial to reduce μ properly in order to achieve maximum efficiency of the continuation method. From our experience, μ should be reduced rapidly when the iterate is very far away from the solution, slowly when closer to a solution of the LCP, and then very fast when very close to the solution. Based on our experience, it seems proper to choose μ according to the following rule:

1. Let $u = err(\mathbf{x}^k)^2/n$, and choose μ^k according to the following formula:

$$\mu^k = \begin{cases} \sqrt{u} & \text{if } u \geq 1, \\ u & \text{otherwise}; \end{cases}$$

2. if $\mu^k < 10^{-10}$, set $\mu^k = 10^{-10}$;
3. if $\mu^k > \mu^{k-1}$, set $\mu^k = \mu^{k-1}$;
4. if $\|\mathbf{J}(\mathbf{x}^k, \mu^{k-1})\| < 10^{-4}$, set $\mu^k = 10^{-2}\mu^k$.

The rules are executed following the order listed above. The basic formula in Rule 1 has the property that when $err(\mathbf{x})$ is very large, μ decreases very fast; as $err(\mathbf{x})$ gets smaller, μ rapidly approaches zero. Rule 2 takes into consideration numerical stability. When μ is very small and thus, the matrix to be inverted is very ill conditioned, μ is reduced carefully and relatively slowly. Rules 3 and 4 make sure that μ decreases monotonically, and therefore, the algorithm converges.

In all the tests reported in the next section, the following parameter values are used: $\beta = 0.5$, $\sigma = 0.1$, $s = 1.0$. The initial vector $\mathbf{x}^0$ is taken to be the zero vector.

The initial μ^0 is taken to be $\|\mathbf{q}\|/n$. The algorithm terminates when $err(\mathbf{x}) \leq \epsilon = 10^{-6}$. The continuation algorithm is implemented in Fortran-77 on an Apollo DN4500 workstation. The iteration number reported is the number of linear equations solved, and all the CPU times reported are in seconds using Apollo system routine *get – cput* and excluding input/output time.

5. Numerical experiments.

In what follows, the continuation method is compared against both the Polynomial Path Following (PPF) algorithm implemented by Mizuno, Yoshise, and Kikuchi [20] and Lemke's algorithm implemented by Tomlin [23], which is available from the Systems Optimization Laboratory at Stanford University.

The test problems are of two categories: constructed problems and randomly generated problems. The constructed problems include two problems for which Lemke's algorithm is known to run in exponential time (see [21, Chap. 6]) and an example constructed by Pang (see [21, Chap. 8]) to test SOR methods.

The randomly generated problems are constructed as follows. First, an $n \times m$ $(n \geq m)$ matrix $\mathbf{A}$ is randomly generated with nonzero entries uniformly distributed in some interval $[a, b]$. The matrix $\mathbf{M}$ in the LCP is then computed as

$$\mathbf{M} = \mathbf{A}\mathbf{A}^T.$$

The rank of $\mathbf{M}$ is at most m and very likely m due to the fact that the elements of $\mathbf{A}$ are randomly generated, and it is almost impossible that two rows of $\mathbf{A}$ are dependent. Therefore, by varying m, one can generate matrices $\mathbf{M}$ of various ranks. In general, the density of $\mathbf{M}$ is positively correlated with that of $\mathbf{A}$, which can be controlled when it is generated. The vectors $\mathbf{q}$ are generated in two different ways. When comparing against PPF, $\mathbf{q}$ is generated as a random vector whose elements are uniformly distributed within some interval $[a, b]$; this is the case reported in [20]. In order to control the number of nonzero $\mathbf{x}$'s in the solution, $\mathbf{q}$ can also be generated in a reverse manner: first, a solution vector $(\mathbf{x}, \mathbf{w})$ is randomly generated that satisfies the requirements for an LCP solution, and $\mathbf{q}$ is then calculated by

$$\mathbf{q} = \mathbf{w} - \mathbf{M}\mathbf{x}.$$

5.1. Comparisons based on constructed problems.

The first example consists of an n-vector $\mathbf{q}$, which has each component equal to minus one and the following matrix $\mathbf{M}$:

$$\mathbf{M} = \begin{pmatrix} 1 & 2 & 2 & \dots & 2 \\ 0 & 1 & 2 & \dots & 2 \\ 0 & 0 & 1 & \dots & 2 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

The second example has the same vector $\mathbf{q}$ but a different matrix $\mathbf{M}$:

$$\mathbf{M} = \begin{pmatrix} 1 & 2 & 2 & \dots & 2 \\ 2 & 5 & 6 & \dots & 6 \\ 2 & 6 & 9 & \dots & 10 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 2 & 6 & 10 & \dots & 4(n-1)+1 \end{pmatrix}.$$

Lemke's algorithm is known to run in exponential time for these two examples. Comparisons between the continuation method and the practical version of the PPF algorithm are listed in Table 1. Only iteration numbers are reported since these two algorithms were run on different machines. The iteration numbers of the PPF algorithm are the best results chosen from [20]. The table shows that the continuation method outperforms the PPF algorithm by a large margin. For all the test problems with various dimensions, the continuation method takes only four to five iterations. In contrast, the PPF algorithm takes a fairly large number of iterations to converge, and the iteration number increases with the problem size.

TABLE 1
Iterations for the exponential examples.

n	PPF Algorithm		Continuation Method	
	Example 1	Example 2	Example 1	Example 2
8	28	32	5	5
16	29	34	5	4
32	30	36	4	4
64	32	38	4	4
128	33	39	4	4

The third example was constructed in order to test various SOR methods. The matrix $\mathbf{M}$ has the following form:

$$\mathbf{M} = \begin{pmatrix} 6 & -4 & 1 & 0 & 0 & 0 & \ldots & 0 \\ -4 & 6 & -4 & 1 & 0 & 0 & \ldots & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 & \ldots & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 & -4 & 6 \end{pmatrix}.$$

Pang [22] tried to solve this class of LCPs for $n = 100$ using various iterative SOR methods and found that convergence was not obtained even after thousands of iterations, while Lemke's method solves these problems very well. The comparison between the continuation method and Lemke's method is listed in Table 2. In our tests, the vector $\mathbf{q}$ is generated in the reverse manner, as described above, where the percentage of nonzeros in the solution is around 50%. The results of the continuation method are very impressive; for problems of various dimensions, the continuation method outperforms Lemke's algorithm and as the problem size increases, the relative performance of the continuation method becomes better.

In summary, the continuation method outperforms both the PPF algorithm and Lemke's method for the above constructed problems. All these problems are very well structured and thus, Newton's method takes only one iteration to reach to the solution of PLCP(μ) for each μ.

5.2. Comparisons based on randomly generated problems. Unlike the linear programming literature, there exist very few standard test problems for LCPs. Therefore, we have to rely on randomly generated problems to test the performance of LCP algorithms.

5.2.1. Comparison against the PPF algorithm. As shown in Table 3, the implementation of the PPF algorithm detailed in [20] does not compare well against the proposed continuation method. The test problems are generated according to the

TABLE 2
Results for Pang's Example

n		% of $x_i > 0$	Lemke's Algorithm Iter.	Lemke's Algorithm CPU sec	Continuation Method Iter.	Continuation Method CPU sec
200	Max.		115	5.76	12	2.86
	Avg.	56.1	112.6	5.634	11.2	2.646
	Min.		109	5.54	11	2.57
300	Max.		158	14.05	11	3.87
	Avg.	49.3	147.8	11.804	10.4	3.674
	Min.		130	9.66	10	3.59
400	Max.		209	24.20	12	5.58
	Avg.	50.4	202.2	23.524	11.2	5.198
	Min.		193	23.00	11	5.14
500	Max.		259	44.47	13	7.42
	Avg.	49.8	249.8	42.798	11.8	6.812
	Min.		242	41.44	11	6.31
600	Max.		307	72.17	13	9.12
	Avg.	49.1	295.6	64.114	12.0	8.256
	Min.		283	58.39	11	7.629

TABLE 3
Iterations for randomly generated problems.

n		PPF Algorithm	Continuation Method
8	Max.	30	8
	Avg.	29.3	6.3
	Min.	27	4
16	Max.	33	7
	Avg.	32.1	6.7
	Min.	31	6
32	Max.	35	10
	Avg.	33.5	7.3
	Min.	32	6
64	Max.	36	9
	Avg.	35.5	7.8
	Min.	35	7
128	Max.	–	10
	Avg.	37	8.8
	Min.	–	8

method reported in [20], where $\mathbf{M} = \mathbf{A} \times \mathbf{A}^T$ is a 100% dense positive semidefinite matrix. The entries of $\mathbf{A}$ and $\mathbf{q}$ are randomly distributed in $[-1, +1]$. Similarly, ten examples were run for problems of dimension 8, 16, 32, and 64. For the problem of dimension 128, only one example was run for the PPF algorithm in [20], while in our test, five examples were run for the continuation method. For the same reasons as given in §5.1, only iteration numbers are reported. The iteration numbers for the PPF algorithm are the best of all the implementations reported in [20]. As one can see from Table 3, the continuation method takes far fewer iterations to converge for problems of all sizes. Since both the PPF algorithm and the continuation method solve the same type of linear equations at each iteration as shown by Proposition 3.10, the overall performance of the continuation method should be much better than the PPF algorithm.

5.2.2. Factors that affect relative performance. We now compare the continuation method against Lemke's algorithm based on randomly generated problems. The following factors could affect their relative performances:

1. Density of **M**: A sparse LCP is favored by both algorithms since sparse linear equation solution techniques are used in both cases. We are interested in the relative performance of these two algorithms as the density of **M** changes.

2. n, the dimension of **M**: As n increases, both algorithms take a longer time to solve the LCP. We are interested in the relative performance as n increases.

3. Rank of **M**: As the rank of **M** gets lower, the matrix to be inverted in the continuation method becomes more ill conditioned. As a result, the solutions of the linear equations are less accurate and the continuation method should take more iterations to converge.

4. The number of nonzeros in the solution and signs of the **q** vector: These two factors are correlated with each other. If there are many negative entries in the **q** vector, it is more likely that the solution will have more nonzero **x**'s and vice versa. This factor does not seem to have any direct effect on the continuation method. However, it dramatically affects the performance of Lemke's algorithm as well as other algorithms, such as the damped Newton method by Harker and Pang [7]. If there are many nonzeros in the LCP solution, the base update of Lemke's algorithm has to invert matrices of larger dimension and thus, the algorithm may take more time to converge.

The above four factors will be tested separately in the following four subsections. While the effects of one factor are tested, other factors remain unchanged. For each problem instance, five random examples were generated unless explicitly stated otherwise. The average, best case, and worst case of these five examples are reported herein.

5.2.3. Effects of the number of nonzeros in the solution. This factor is tested first since it seems to have the largest impact on the performance of the algorithms under investigation. The average density of the examples is listed as an extra argument since it is difficult to generate series of random matrices with identical density. The parameters of the problem are chosen as follows:

- Dimension of **M**: 300
- Rank of **M**: 300
- Range of entries in **M**: $[-5, +5]$
- Range of nonzero solutions: $[+10, +20]$

The results are listed in Table 4, and the relative performance of Lemke's algorithm versus the continuation method is shown in Fig. 1, where the vertical axis is the ratio of their CPU times and is plotted using a logarithmic scale (base 2). As we can see in Table 4, the performance of the continuation method is relatively insensitive to this factor, while that of Lemke's algorithm changes dramatically. Lemke's algorithm dominates the continuation method when there are fewer nonzeros in the solution, while the continuation method performs better in the opposite situation. For the test problems we ran, the two algorithms are tied when the percentage nonzeros in the solution is between 60% and 70%. However, as we shall see in §5.2.6, the break-even point becomes smaller when matrix **M** in the LCP is sparser and larger. Therefore, the continuation method could be used as an alternative method to solve LCPs when many nonzeros are expected in the solutions.

5.2.4. Effects of the density of M. Like the interior-point algorithm for linear programming, the continuation algorithm is competitive with Lemke's method only when the matrix **M** is sparse. The results of increasing the density are listed in Table 5, and the relative performance is shown in Figure 2, where the vertical axis is the

TABLE 4
Effects of the number of nonzeros in the solution.

	% of $x_i > 0$	Density of M	Lemke's Algorithm Iter.	Lemke's Algorithm CPU sec	Continuation Method Iter.	Continuation Method CPU sec
Max.			160	28.06	11	97.75
Avg.	28.7	5.03	146.8	24.51	9.8	85.06
Min.			126	21.21	9	74.89
Max.			204	46.53	10	98.09
Avg.	40.7	5.25	192.2	41.36	9.6	87.33
Min.			174	34.71	8	75.46
Max.			244	64.08	12	105.5
Avg.	50.7	5.23	227.0	62.51	10.4	100.2
Min.			202	48.70	10	87.02
Max.			255	88.47	11	103.5
Avg.	60.2	5.18	242.4	81.44	10.6	94.32
Min.			221	74.38	10	74.86
Max.			279	126.5	12	125.5
Avg.	68.2	5.21	267.0	112.4	10.2	97.45
Min.			253	93.30	9	81.91
Max.			302	202.3	11	124.1
Avg.	80.1	5.29	284.0	151.6	10.2	95.87
Min.			266	124.4	9	82.82

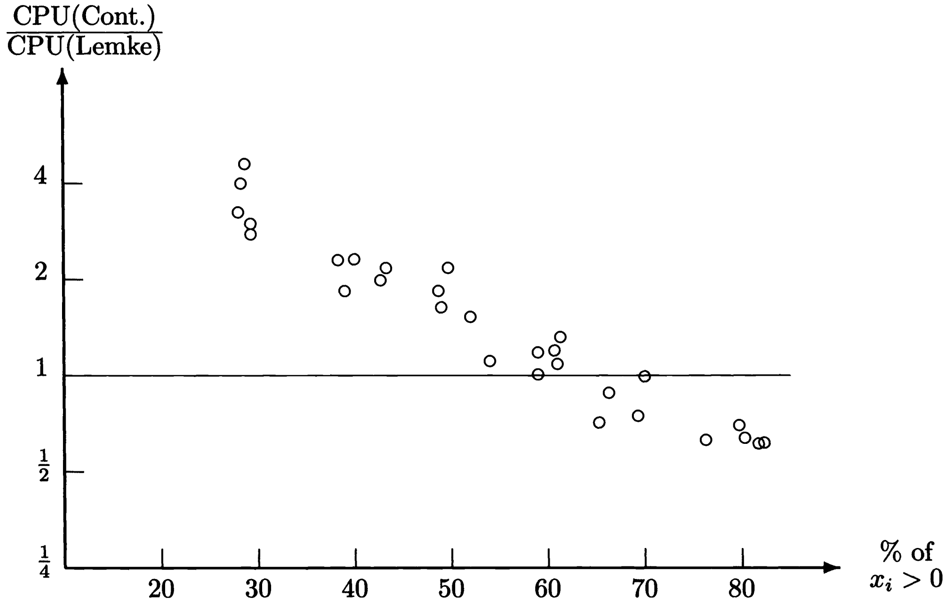


FIG. 1. *Effects of number of nonzeros in the solution.*

ratio of their CPU times. The average percentage of nonzeros in the solution is listed as an extra argument. The parameters of the test problems are chosen as follows:

- Dimension of **M**: 300
- Rank of **M**: 300
- Range of entries in **M**: $[-5, +5]$

- Range of nonzero solutions: $[+10, +20]$

It is clear from Table 5 and Fig. 2 that the performance of the continuation method is affected more by the density of **M** than is Lemke's algorithm. The sparser the matrix **M**, the better the relative performance of the continuation method against Lemke's algorithm. One thing that does not emerge in Table 5 is the fact that the continuation method is less stable when the density is very low. In our tests, the continuation method failed to converge several times when **M** is very sparse. However, this problem vanishes as the dimension of **M** gets larger. The reason for this is not fully understood.

TABLE 5
Effects of the density of **M**.

	Density of M	% of $x_i > 0$	Lemke's Algorithm Iter.	Lemke's Algorithm CPU sec	Continuation Method Iter.	Continuation Method CPU sec
Max.			229	55.11	11	59.82
Avg.	3.98	50.2	207.0	46.63	9.8	52.36
Min.			194	41.50	9	45.31
Max.			215	56.93	9	81.51
Avg.	5.08	49.9	205.2	52.79	8.6	77.07
Min.			190	47.94	8	69.52
Max.			251	96.66	9	119.1
Avg.	6.04	50.9	235.4	80.70	8.8	108.7
Min.			223	70.42	8	93.94
Max.			240	90.29	10	140.8
Avg.	6.88	50.3	229.2	82.10	9	134.8
Min.			219	77.11	8	128.5
Max.			227	94.48	9	176.4
Avg.	8.05	50.2	222.6	88.37	8.8	167.0
Min.			213	81.29	8	147.1
Max.			249	120.3	9	198.3
Avg.	8.93	47.3	227.0	96.62	8.8	179.3
Min.			212	81.02	8	165.8

5.2.5. Effects of the rank of M. The results of the test dealing with the rank of **M** are listed in Table 6, and the relative performance is shown in Fig. 3, where the vertical axis is the same as before. The average density and percentage of nonzeros in the solution are listed as extra arguments. The parameters of the test problems are chosen as follows:

- Dimension of **M**: 300
- Range of entries in **M**: $[-5, +5]$
- Range of nonzero solutions: $[+10, +20]$

As one can see from Table 6, the continuation method takes more iterations to converge when **M** is of lower rank, while Lemke's algorithm is not sensitive to this factor. In our experiments, the continuation method failed to converge several times when the rank of **M** is very low. It turns out that whenever the method fails, the solution of linear equations are very inaccurate because of ill conditioning and thus, the line search cannot find a descent direction. Similar to the test of density, this is less of a problem when the dimension of **M** becomes larger.

5.2.6. Effects of the dimension of M. The results of increasing the problem dimension are listed in Table 7, where problems of larger sizes are tested. The relative performance is shown in Fig. 4, where the vertical axis is the same as before. Only one example of dimension 1500 is run because of excessive computational time. The

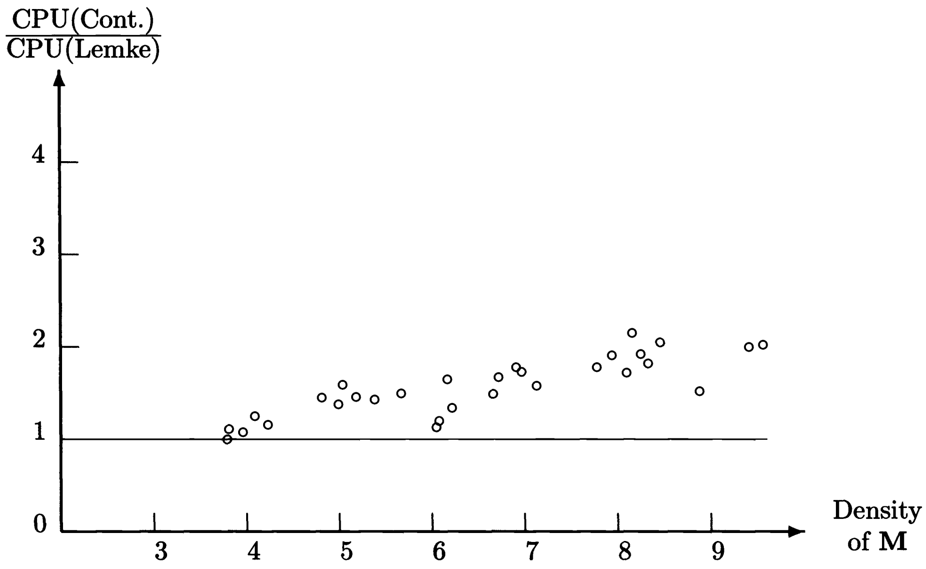


FIG. 2. *Effects of the sparsity of* **M**.

TABLE 6
Effects of the rank of **M**.

Rank		Density	% of	Lemke's Algorithm		Continuation Method	
of **M**		of **M** (%)	$x_i > 0$	Iter.	CPU sec	Iter.	CPU sec
	Max.			251	79.20	17	200.0
200	Avg.	6.00	48.6	231.8	67.61	14.6	147.5
	Min.			198	52.34	11	113.0
	Max.			248	88.12	17	180.2
220	Avg.	6.14	48.3	228.2	71.38	13.0	146.4
	Min.			211	60.46	10	105.1
	Max.			249	91.90	13	143.7
240	Avg.	6.01	51.7	235.6	78.57	11.4	128.6
	Min.			217	64.66	9	100.6
	Max.			248	82.14	10	118.3
260	Avg.	6.04	50.6	235.2	75.43	9.6	109.6
	Min.			227	70.57	9	96.39
	Max.			227	78.91	11	116.4
280	Avg.	5.88	50.3	225.4	73.93	9.2	100.1
	Min.			142	65.00	8	89.36

average density of **M** and the percentage of nonzeros in the solution are listed as extra arguments in Table 7. The parameters of the test problems are as follows:

- Rank of **M**: $0.9n$
- Range of entries in **M**: $[-5, +5]$
- Range of nonzero solutions: $[+10, +20]$

A very special feature of the continuation method is that the iteration number is very insensitive to the problem size, as one can see in Table 7, although the reason for this is not quite understood. For problems of the same density and percentage of nonzeros in the solution, the relative performance of the continuation method becomes better as the problem size increases. For the test problem we generated, Lemke's algorithm is as fast as the continuation method when the dimension of **M** is 600. However,

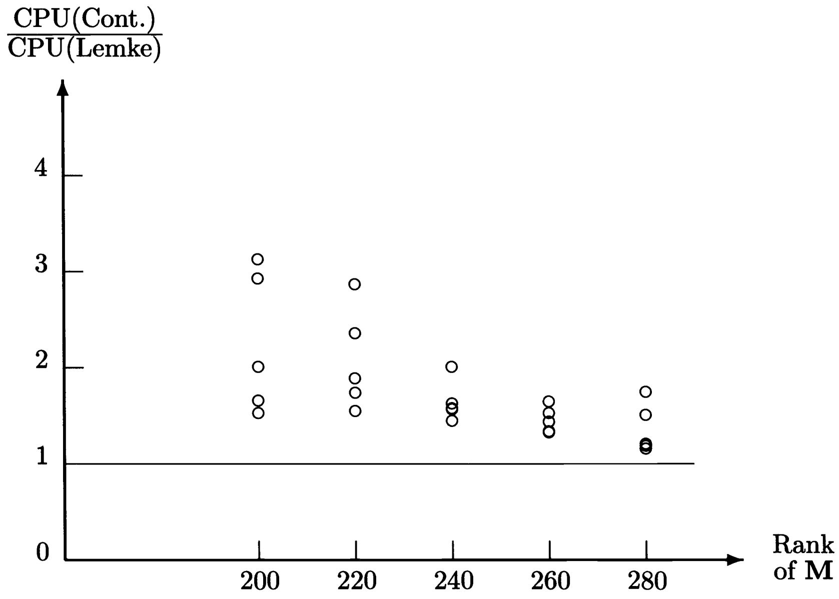


FIG. 3. *Effects of the rank of* **M**.

the continuation method is much faster when the dimension of **M** increases to 1000 or larger. Therefore, one can imagine that when the problem size is large enough, and matrix **M** is reasonably sparse, the continuation method can be competitive with Lemke's algorithm even if the LCP has less than 50% nonzeros in the solution.

TABLE 7
Effects of the dimension of **M**.

n		Density of M	% of $x_i > 0$	Lemke's Algorithm		Continuation Method	
				Iter.	CPU sec	Iter.	CPU sec
	Max.			503	849.8	11	858.0
600	Avg.	3.04	60.4	487.8	793.9	10.0	768.4
	Min.			472	727.7	9	677.9
	Max.			685	3760	11	3044
800	Avg.	2.99	61.5	674.6	3161.7	10.0	2801.7
	Min.			476	2519	9	2345
	Max.			948	11165	10	7022
1000	Avg.	3.03	60.5	881.2	9356.0	9.6	6672.3
	Min.			835	7507	9	6331
	Max.			1095	21652	11	15351
1200	Avg.	3.00	60.2	1053.0	19661.2	9.8	13906.0
	Min.			1009	16077	9	12679
	Max.						
1500	Avg.	2.99	59.1	1286	54947	11	36701
	Min.						

6. Conclusion and future research. This paper modifies the PPF for the monotone LCP and extends the algorithm to solve LCPs defined by a P_0-matrix. The most attractive property of the continuation algorithm is its flexibility compared with the PPF. The initial and intermediate iterates do not have to stay interior although the path followed by our continuation method, or the path of centers, is the same as that of

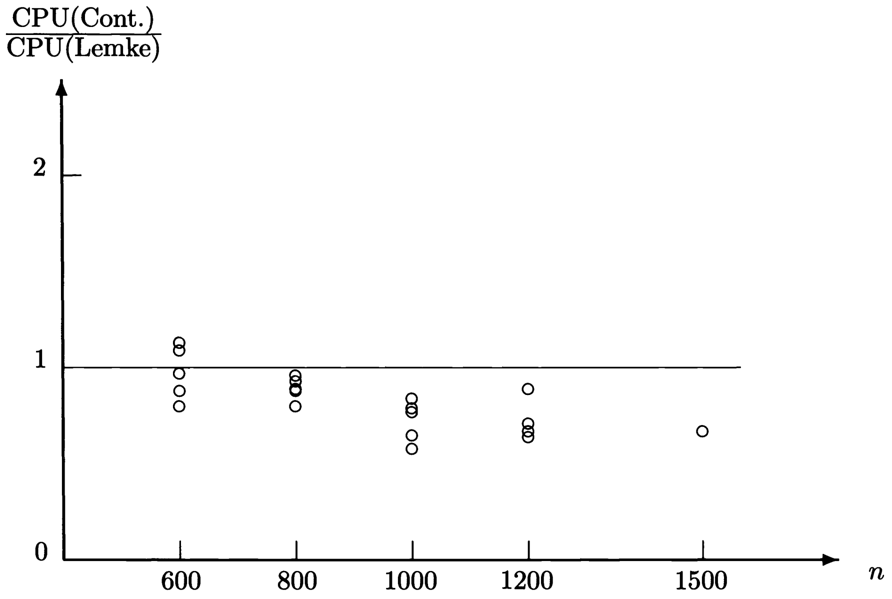


FIG. 4. *Effects of the dimension of* **M**.

the PPF and is contained in the interior. The step size in the continuation algorithm can be adjusted with more freedom, and a line search can easily be incorporated. The initial continuation parameter μ can be set to be very small (except, of course, for numerical considerations) and can be reduced quickly in order to achieve more rapid convergence.

The numerical experiments presented in this paper show that the proposed continuation method is a viable algorithm for LCPs with positive semidefinite matrices. The algorithm outperforms Lemke's method for those LCPs with large, sparse, and high-rank **M** and have relatively more nonzeros in the solution. Since the continuation method takes a very small number of iterations to converge, it has great potential for very large LCPs. More effort should be spent on improving the speed of solving the resulting linear equations, which takes at least 95% of the CPU time. The current code uses an unmodified version of SPARSPAK for solving the linear equations. To make the code more competitive, SPARSPAK should be modified so that some of the interface programs are deleted. The improved code should also take care of numerical instability; that is, to improve the accuracy of the solution of the linear equations.

Future research will be directed toward the following:

- computational complexity analysis of the continuation method,
- more rigorous comparison between the PPF algorithm and the continuation method proposed in this paper,
- theoretical evidence on reducing μ optimally, since this procedure is crucial for the continuation method.

Future research will also be devoted to efficiently solving nonlinear complementarity problems, nonlinear programs [3], and linear programs as well [4].

Acknowledgments. This work was supported by the National Science Foundation Presidential Young Investigator Award ECE-8552773 and by a grant from the Burlington Northern Railroad. The authors gratefully acknowledge the comments and encouragement of Jong-Shi Pang. We are also indebted to two referees for their helpful corrections and suggestions that lead to the improvement of the manuscript.

REFERENCES

[1] M. AGANAGIC AND R. W. COTTLE, *A note on Q-matrices*, Math. Programming, 16 (1979), pp. 374–377.

[2] D.P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[3] B. CHEN AND P. T. HARKER, *A Non-Interior-Point Continuation Method for Monotone Variational Inequalities*, Working Paper 90-10-02, Decision Sciences Department, Wharton School, University of Pennsylvania, Philadelphia, PA, 1990; Math Programming, to appear.

[4] ———, *A Noninterior Continuation Method for Quadratic and Linear Programming*, Working Paper 90-11-05, Decision Sciences Department, Wharton School, University of Pennsylvania, Philadelphia, PA, 1990; SIAM J. Optim., 3 (1993), pp. 503–515.

[5] E. C. H. CHU, J. W. H. LIU, J. A. GEORGE, AND E. G. Y. NG, *SPARSPAK: Waterloo Sparse Matrix Package, User's Guide for SPARSPAK-A*, Research Report CS-84-36, Department of Computer Science, University of Waterloo, Ontario, Canada, 1984.

[6] M. FIEDLER AND V. PTAK, *Some generalizations of positive definiteness and monotonicity*, Numer. Math., 9 (1966), pp. 163–172.

[7] P. T. HARKER AND J. S. PANG, *A damped-Newton method for the linear complementarity problem*, AMS Lectures in Applied Mathematics, 26 (1990), pp. 265–283.

[8] M. KOJIMA, N. MEGIDDO, AND T. NOMA, *Homotopy continuation methods for complementarity problems*, Research Report RJ 6638, IBM Almaden Research Center, San Jose, CA, 1989.

[9] M. KOJIMA, N. MEGIDDO, AND Y. YE, *An Interior Point Potential Reduction Algorithm for the Linear Complementarity Problem*, Research Report RJ 6486, IBM Almaden Research Center, San Jose, CA, 1988.

[10] M. KOJIMA, N. MEGIDDO, T. NOMA, AND A. YOSHISE, *A unified approach to interior point algorithms for linear complementarity problem*, Lecture Notes in Comput. Sci. 538, Springer-Verlag, New York, 1991.

[11] M. KOJIMA, S. MIZUNO, AND T. NOMA, *A new continuation method for complementarity problems with uniform P-functions*, Math. Programming, 43 (1989) pp. 107–113.

[12] ———, *Limiting behavior of trajectories generated by a continuation method for monotone complementarity problems*, Math. Oper. Res., 15 (1990), pp. 662–675.

[13] M. KOJIMA, S. MIZUNO, AND A. YOSHISE, *An $O(\sqrt{n}L)$ Iteration Potential Reduction Algorithm for Linear Complementarity Problems*, Research Report on Information Sciences B-217, Tokyo Institute of Technology, Tokyo, Japan, 1988.

[14] ———, *A polynomial-time algorithm for a class of linear complementarity problems*, Math. Programming, 44 (1989), pp. 1–26.

[15] J. W. H. LIU, *Modification of the minimum-degree algorithm by multiple elimination*, ACM Trans. Math. Software, 11 (1985), pp. 141–153.

[16] D. G. LUENBERGER, *Linear and Nonlinear Programming*, 2nd ed., Addison-Wesley, Reading, MA, 1984.

[17] I. J. LUSTIG, R. E. MARSTEN, AND D. F. SHANNO, *Computational Experience with a Primal-Dual Interior Point Method for Linear Programming*, Tech. Report J-89-11, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 1989.

[18] R. MATHIAS AND J. S. PANG, *Error bounds for the linear complementarity problem with a P-matrix*, Linear Algebra Appl., 132 (1990), pp. 123–136.

[19] S. MIZUNO, *A New Polynomial Time Method for a Linear Complementarity Problem*, Tech. Report No. 16, Department of Industrial Engineering and Management, Tokyo Institute of Technology, Tokyo, Japan, 1989.

[20] S. MIZUNO, A. YOSHISE, AND T. KIKUCHI, *Practical Polynomial Time Algorithms for Linear Complementarity Problems*, Tech. Report No. 13, Department of Industrial Engineering and Management, Tokyo Institute of Technology, Tokyo, Japan, 1989.

[21] K. G. MURTY, *Linear Complementarity, Linear and Nonlinear Programming*, Helderman-Verlag, Berlin, 1988.

[22] J. S. PANG, private communication, 1990.

[23] J. A. TOMLIN, *USERS GUIDE FOR LCPL: A Program for Solving Linear Complementarity Problems by Lemke's Algorithm*, Tech. Report SOL 76-16, Department of Operations Research, Stanford University, Stanford, CA, 1976.

[24] L. T. Watson, J. P. Bixler, and A. B. Poore, *Continuous Homotopies for the Linear Complementarity Problems*, Working Paper, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1987.

[25] Y. Ye, *A Further Result on the Potential Reduction Algorithm for the P-Matrix Linear Complementarity Problem*, Manuscript, Department of Management Sciences, University of Iowa, Iowa City, IA, 1988.

[26] ———, *Bimatrix Equilibrium Points and Potential Functions*, Working Paper 88-16, Department of Management Sciences, University of Iowa, Iowa City, IA, 1988.

[27] Y. Ye and P. Pardalos, *A Class of Linear Complementarity Problems Solvable in Polynomial Time*, Working Paper, Department of Management Sciences, University of Iowa, Iowa City, IA, 1989.

[28] W. I. Zangwill and C. B. Garcia, *Pathways to Solutions, Fixed Points, and Equilibria*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

SIAM J. MATRIX ANAL. APPL.
Vol. 14, No. 4, p. 1191, October 1993

022

ERRATUM: A LOOK-AHEAD LEVINSON ALGORITHM FOR INDEFINITE TOEPLITZ SYSTEMS*

TONY F. CHAN† AND PER CHRISTIAN HANSEN‡

Freund and Zha [2] pointed out that our algorithm in [1] can break down in the rare situation where two consecutive block steps are taken. Motivated by the algorithm in [2], we show here how this error in our algorithm can be corrected within the formulation used in [1] by changing Theorem 3 of [1] (and the corresponding step in the algorithm) as follows (the proof is straightforward and is omitted here).

THEOREM 3. *In case of two consecutive extended Levinson steps with $k = k' + p'$, Theorem 1 can be used with the vector $\mathbf{s}_k$ given by*

$$\mathbf{s}_k \equiv \begin{pmatrix} E_{k'} Y_{p'} \\ I_{p'} \end{pmatrix} \left(\Gamma_{p'}^{(k')} \right)^{-1} \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix},$$

where the matrices $\Gamma_{p'}^{(k')}$ and $Y_{p'}$ are associated with the previous block step.

REFERENCES

[1] T. F. CHAN AND P. C. HANSEN, *A look-ahead Levinson algorithm for indefinite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 490–506.

[2] R. FREUND AND H. ZHA, *Formally biorthogonal polynomials and a look-ahead Levinson algorithm for general Toeplitz systems*, Report CS-92-23, Dept. of Computer Science, Pennsylvania State University, University Park, PA, October 1992; Linear Algebra Appl., to appear.

* Received by the editors May 27, 1993; accepted for publication May 27, 1993.

† Dept. of Mathematics, University of California, 405 Hilgard Avenue, Los Angeles, California 90024-1555 (chan@math.ucla.edu).

‡ UNI•C (Danish Computing Center for Research and Education), Building 305, Technical University of Denmark, DK-2800 Lyngby, Denmark (unipch@wuli.uni-c.dk).