SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 1, January 1988

EQUALITY CLASSES OF MATRICES*

DANIEL HERSHKOWITZ† AND HANS SCHNEIDER†‡

Abstract. Recent results of Neumaier for irreducible matrices on the equality case of a classical matrix inequality due to Ostrowski are generalized to general matrices. Several graph and number theoretic concepts are employed in the proof of various further results.

Key words. irreducible matrix, Frobenius normal form, diagonal similiarity, equality class, twist, cycle product, access cover

AMS(MOS) subject classifications. 15, 05

1. Introduction. Let A be a complex $n \times n$ matrix and define the absolute value matrix $B = |A|$ of A by $b_{ij} = |a_{ij}|$, $i, j = 1, \cdots, n$. Let $\rho(A)$ be the spectral radius of A.

Let $\mathcal{U}$ be the set of all complex matrices A such that $\rho(|A|) < 1$. In [7] Ostrowski proves the now very well known result that, for $A \in \mathcal{U}$,

$$|(I-A)^{-1}| \leqq (I-|A|)^{-1}, \tag{1.1}$$

where the inequality is entrywise.

In [6] Neumaier shows that for $A \in \mathcal{J}$, the set of $n \times n$ irreducible matrices $A \in \mathcal{U}$,

$$|(I-A)^{-1}| = (I-|A|)^{-1}, \tag{1.2}$$

if and only if

(1.3) all circuit products of A are positive.

It is well known ([2], [3]) that for irreducible A, (1.3) is equivalent to

(1.4) A is diagonally similar to $|A|$, i.e., there exists a diagonal matrix X such that $A = X|A|X^{-1}$.

Neumaier also shows in [6] that the condition

$$|(I-A^{-1})|_{ij} = (I-|A|^{-1})_{ij}, \quad \text{for some } i,j, \quad 1 \leqq i,j \leqq n, \tag{1.5}$$

which is apparently weaker than (1.2), is in fact equivalent to (1.2)–(1.4) for $A \in \mathcal{J}$. (We have stated special cases of the results of Ostrowski and Neumaier, from which, however, the general theorems may easily be derived.)

In this paper we generalize Neumaier's results in various directions. We consider the equality (1.2) for general $A \in \mathcal{U}$, omitting the requirement of irreducibility. We use the concept of two-twisted chain of the graph $G(A)$ of A, which was defined in [5] (see also § 2 of this paper). Intuitively, a chain in a directed graph is obtained by putting a pointer at a vertex and moving it either in the direction or against the direction of a connected sequence of arcs to another vertex. Each change in direction is a twist. A two-twisted chain (e.g., cycle) is a chain with at most two twists. Thus, a circuit (directed

* Received by the editors May 19, 1986; accepted for publication October 27, 1987. This work was supported by the United States–Israel Binational Science Foundation under joint grant 85-00153.

† Mathematics Department, University of Wisconsin, Madison, Wisconsin 53706 and Mathematics Department, Technion-Israel Institute of Technology, Haifa 32000, Israel. The first author is permanently located at the Technion and the second at the University of Wisconsin.

‡ The work of this author was supported in part by National Science Foundation grants DMS-8320189 and DMS-8521521, Office of Naval Research grant N00014-85-K-1613, and the Lady Davis Foundation of Israel.

cycle) is a special case of a two-twisted cycle. We show that, for $A \in \mathcal{U}$, condition (1.2) is equivalent to

(1.6) all cycle products of A corresponding to two-twisted cycles are positive

(and other conditions). This generalizes (1.3).

If C is an $s \times s$ matrix and A is an $n \times n$ matrix, where $s \leqq n$, we generalize both the Kronecker and Hadamard products in [4] by defining the $n \times n$ matrix $C \times\times A$, see also § 3. Thus, if A is partitioned into s^2 matrices A_{ij}, $i, j = 1, \cdots, s$, then $C \times\times A$ is the matrix whose blocks are $c_{ij}A_{ij}$, $i, j = 1, \cdots, s$. Here we show that if $A \in \mathcal{U}$ is in Frobenius normal form then A satisfies (1.2) if and only if

(1.7) A is diagonally similar to $C \times\times |A|$, where C is an upper triangular $s \times s$ matrix ($s \leqq n$) such that $|c_{ij}|$ is 1 or 0, c_{ii} is 1 or 0, $i,j = 1, \cdots, s$, and zC satisfies (1.2) for $0 < z < 1$.

This generalizes (1.4).

We also generalize (1.5) by defining the concept of a $G(A)$-access cover, see also § 2. A subset Γ of $\langle n \rangle \times \langle n \rangle$, where $\langle n \rangle = \{1, \cdots, n\}$, is a $G(A)$-access cover if for each $(i, j) \in \langle n \rangle \times \langle n \rangle$ there is an $(h, k) \in \Gamma$ such that h has access to i in $G(A)$ and j has access to k in $G(A)$. We observe that $\{(i, j)\}$ is a $G(A)$-access cover for all $(i, j) \in \langle n \rangle \times \langle n \rangle$ if and only if A is irreducible (or equivalently, $G(A)$ is strongly connected). Thus, if Γ is a $G(A)$-access cover and $A \in \mathcal{U}$, then (1.2) is equivalent to

$$|(I-A)^{-1}|_{ij} = (I-|A|)^{-1}_{ij} \quad \text{for } (i,j) \in \Gamma. \tag{1.8}$$

The results above may be found as part of Theorem 5.14.

It is easily seen that (1.2) is equivalent to

$$\left| \sum_{s \in N} A^s \right| = \sum_{s \in N} |A|^s \tag{1.9}$$

for $A \in \mathcal{U}$, where N is the set of natural numbers. Since, for all subsets S of N,

$$\left| \sum_{s \in S} A^s \right| \leqq \sum_{s \in S} |A|^s, \tag{1.10}$$

it is natural to define Equ ($\mathcal{A}$, Γ, S) to be the set of all $A \in \mathcal{A}$ such that

$$\left| \sum_{s \in S} A^s \right| = \sum_{s \in S} |A|^s \quad \text{for } (i,j) \in \Gamma, \tag{1.11}$$

where $\mathcal{A} \subseteq \mathcal{U}$, $\Gamma \subseteq \langle n \rangle \times \langle n \rangle$ and $S \subseteq N$.

The equivalences stated above, and others, are stated in terms of Equ ($\mathcal{U}$, Γ, N). It is clear that Equ ($\mathcal{A}$, Γ, S) $\supseteq$ Equ ($\mathcal{A}$, Γ, N) for $S \subseteq N$. We therefore call a subset S of N ($\mathcal{A}$, Γ)-sufficient if Equ ($\mathcal{A}$, Γ, S) = Equ ($\mathcal{A}$, Γ, N).

We give conditions equivalent to ($\mathcal{J}$, $\langle n \rangle \times \langle n \rangle$)-sufficiency and ($\mathcal{U}$, $\langle n \rangle \times \langle n \rangle$)-sufficiency. The general problem of characterizing ($\mathcal{A}$, Γ)-sufficient sets and minimal ($\mathcal{A}$, Γ)-sufficient sets, for $\mathcal{A} \subseteq \mathcal{U}$ and $\Gamma \subseteq \langle n \rangle \times \langle n \rangle$, is open.

Section 2 contains graph theoretic preliminaries. Section 3 contains preliminaries from combinatorial matrix theory. The basic definitions and results on Equ ($\mathcal{A}$, Γ, S) are collected in § 4. Sections 5 and 6 contain our principal results on Equ ($\mathcal{A}$, Γ, N) and ($\mathcal{J}$, Γ)-sufficient and ($\mathcal{U}$, Γ)-sufficient sets.

2. Graph theoretic definitions and preliminaries.

DEFINITION 2.1. A (simple, directed) *graph* $G = (V, E)$ is a pair of finite sets with $E \subseteq V \times V$. An element of V is called a *vertex* of G, and an element of E is called an *arc* of G. We call $G = (V', E')$ a *subgraph* of G if $V' \subseteq V$ and $E' \subseteq E$.

DEFINITION 2.2. Let G be a graph. A *chain* in G of *length* s from a vertex i_0 to a vertex i_s of G is a sequence

$$\gamma = (i_0, e_1, i_1, e_2, i_2, \cdots, i_{s-1}, e_s, i_s) \tag{2.3}$$

where either $e_p = 1$ and (i_{p-1}, i_p) is an arc of G or $e_p = -1$ and (i_p, i_{p-1}) is an arc of G, $p = 1, \cdots, s$. The arc (i_{p-1}, i_p), $[(i_p, i_{p-1})]$, $1 \leqq p \leqq s$, is said to *lie on* γ if $e_p = 1$ $[e_p = -1]$. The length of a chain γ is denoted by $|\gamma|$. The chain γ is *simple* if the vertices $i_0, \cdots, i_s$ are distinct. The chain γ is *closed* if $i_0 = i_s$, and γ is called a *cycle* if it is closed and the vertices $i_1, \cdots, i_s$ are distinct. A chain given by (2.3) such that $e_1 = \cdots e_s = 1$ is called a *path*. A path that is a cycle is called a *circuit*. A closed chain of form

$$\gamma = (i_0, e_1, i_1, \cdots, i_s, -e_s, i_{s-1}, \cdots, -e_1, i_0)$$

will be called *trivial*. The empty chain $\varnothing$ will be considered a chain of length 0 from any vertex to itself and is defined to be simple. The set $\{i_0, \cdots, i_m\}$ is called the *vertex set* of the chain g given by (2.3).

Thus the empty chain is the only simple circuit.

Intuitively, the chain (i, e, j) is a step from vertex i to vertex j along the arc (i, j) if $e = 1$ and a step from i to j along the arc (j, i) if $e = -1$. We normally write $i \rightarrow j$ or $i \leftarrow j$ in place of (i, e, j) accordingly as $e = 1$ or $e = -1$. For example, $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$ is a circuit and $1 \rightarrow 2 \rightarrow 3 \leftarrow 1$ is a cycle. Note also that as a consequence of the above definition certain chains are cycles that normally are not considered as such, e.g., $1 \rightarrow 2 \leftarrow 1$. It would make no difference to our results to eliminate such cycles from consideration.

DEFINITION 2.4. A vertex i *has access to* a vertex j in a graph G if there is a path from i to j in G and we write $i >\!\!- j$ or $j -\!\!< i$. If U, W are subsets of the vertex set V of G, then the notation $U >\!\!- W$ indicates that every vertex of U has access to every vertex of W.

Observe that a vertex i has access to itself since $\varnothing$ is a path from i to i.

DEFINITION 2.5. A graph G is *strongly connected* if every vertex of G has access to every vertex of G. A subgraph H of G is called a *component* of G if H is a maximal strongly connected subgraph of V, viz. H is strongly connected but no subgraph properly containing H is connected.

DEFINITION 2.6. Let $G = (V, E)$ be a graph and let $(i, j), (h, k) \in V \times V$. Then (i, j) is a *G-access cover* for (h, k) (or (i, j) *G-access covers* (h, k)) if $i >\!\!- h$ and $k >\!\!- j$. Let Γ be a subset of $V \times V$. Then the set of all (h, k) that are G-access covered by elements of Γ will be denoted by $A_G(\Gamma)$. If $\Lambda \subseteq A_G(\Gamma)$, we shall say that Γ is a *G-access cover* for Λ (or that Γ *G*-access covers Λ). If α is a chain in G [G' is a subgraph of G] with vertex set V', then Γ will be called a *G-access cover* for α [G'] if Γ access covers $V' \times V'$. A G-access cover for $V \times V$ will be called a *G-access cover.*

It is easy to show that A_G considered as an operator from the set of subsets of $V \times V$ into itself is a closure operator in the sense of [1, p. 42].

The following lemma is clear:

LEMMA 2.7. *Let $G = (V, E)$ be a graph. Then the following conditions are equivalent*:

(i) *G is strongly connected.*

(ii) *Every nonempty subset of $V \times V$ is a G-access cover.*

(iii) *Every pair $(i, j) \in V \times V$ is a G-access cover.*

Remark 2.8. Let G be a graph and let $H_1, \cdots, H_s$ be the components of G with vertex sets $V_1, \cdots, V_s$, respectively. It is possible to order the components of G so that

$$V_p > - V_q \Rightarrow p < q \quad \text{for } p, q = 1, \cdots, s.$$

DEFINITION 2.9. (i) Let β and γ be the chains $(i_0, e_1, \cdots, e_s, i_s)$ and $(j_0, f_1, \cdots, f_t, j_t)$, respectively. If $i_s = j_0$ we define the *concatenated chain* $\beta\gamma$ by $(i_0, e_1, \cdots, e_j, i_s, f_1, \cdots, f_t, j_t)$. (If $i_s \neq j_0$ then $\beta\gamma$ is not defined.)

(ii) Let α and β be chains. We call α an *extension* (chain) of β (and β a *subchain* of α) if $\beta = \beta_1\beta_2$ and $\alpha = \beta_1\alpha'\beta_2$ where β_1, β_2, and α' are chains (which may be empty). Also, an extension of an extension of β is defined to be an extension of β.

It is easy to see that if α is an extension of β then α and β may be written in the forms $\beta = \beta_1\beta_2 \cdots \beta_p$ and $\alpha = \alpha_0\beta_1\alpha_1 \cdots \beta_p\alpha_p$, where the α_i, $i = 0, \cdots, p$, β_i, $i = 1, \cdots, p$ are chains and α_i, $i = 1, \cdots, p-1$ is closed.

DEFINITION 2.10.

(i) Let γ be the chain given by (2.3). Then the *reverse chain* of γ is defined to be $(i_s, -e_s, i_{s-1}, \cdots, -e_1, i_0)$, and is denoted by γ^*.

(ii) We call e_1 [e_s] the *initial* [*final*] *sign* of γ.

DEFINITION 2.11. Let γ be a chain given by (2.3).

(i) If $e_p \neq e_{p+1}$, $1 \leqq p < s$, then we say that γ *has a twist at* p (or that p is a *twist* of γ). If γ is a closed chain then we allow $p = 0$ and we let $e_0 = e_s$.

(ii) If γ has exactly k twists then γ is said to be *exactly k-twisted* and we put $t(\gamma) = k$.

(iii) If $t(\gamma) \leqq m$ for an integer m then γ is said to be *m-twisted.*

Note that if γ is not closed then $t(\gamma)$ is equal to the number of sign changes in the sequence $e_1, \cdots, e_s$. If γ is closed then $t(\gamma)$ is equal to the number of sign changes in the sequence $e_1, \cdots, e_s, e_1$. Also note that a closed chain in form (2.3) may have a twist at $0, \cdots, s-1$ but not at s.

Observe that a chain [cycle] is 0-twisted if and only if it is a path [circuit] or a reversed path [reversed circuit], and that a closed chain has an even number of twists.

LEMMA 2.12. *Let G be a graph.*

(i) *If α is a chain in G and γ is a subchain of α then*

$$t(\gamma) \leqq t(\alpha) + 1. \tag{2.13}$$

(ii) *If, further, α and γ are closed then*

$$t(\gamma) \leqq t(\alpha). \tag{2.14}$$

Proof. (i) Let $\gamma = \gamma_1 \cdots \gamma_p$ and $\alpha = \alpha_0\gamma_1\alpha_1 \cdots \alpha_{p-1}\gamma_p\alpha_p$. We shall establish a $1-1$ mapping of the set of twists of γ (excluding a possible twist at 0) into the set of twists of α. Suppose that $|\alpha_i| = s_i$, $i = 0, \cdots, p$ and that $|\gamma_i| = t_i$, $i = 1, \cdots, p$. Let $1 \leqq r \leqq t_1 + \cdots + t_p$ and suppose that γ has a twist at r. Then

$$r = t_1 + \cdots + t_i + q$$

where $0 \leqq i < p$ and $1 \leqq q \leqq t_{i+1}$. If $q < t_{i+1}$ then α has a twist at $r + s_0 + \cdots + s_i$. If $q = t_{i+1}$ then $i < p - 1$ (since γ does not have a twist at $t_1 + \cdots + t_p$) and, since the final sign of γ_{i+1} and the initial sign of γ_{i+2} are inequal, it follows that α must have a twist at $r + s_1 + \cdots + s_i + q'$ for some q' satisfying $0 \leqq q_i \leqq s_{i+1}$. This proves the existence of the claimed injection and (i) follows.

(ii) If α and γ are closed, then $t(\alpha)$ and $t(\gamma)$ are both even and (ii) follows from (i). □

3. Definitions and preliminaries in combinatorial matrix theory.

DEFINITION 3.1. Let c be a complex number. The *sign* of c is defined by

$$\operatorname{sgn}(c) = \begin{cases} c/|c|, & \text{if } c \neq 0. \\ 0, & \text{if } c = 0. \end{cases}$$

We call a complex number c a *sign* if $|c|$ is either 0 or 1. If $A \in \mathbb{C}^{nn}$, then we call A a *sign matrix* if a_{ij} is a sign for $i, j = 1, \cdots, n$.

DEFINITION 3.2. Let $A \in \mathbb{C}^{nn}$.

(i) Then $C = |A| \in \mathbb{C}^{nn}$ is defined by $c_{ij} = |a_{ij}|$ for $i, j = 1, \cdots, n$.

(ii) The matrix A is called *nonnegative* ($A \geqq 0$) if $a_{ij} \geqq 0$, $i, j = 1, \cdots, n$.

DEFINITION 3.3. If $A \in \mathbb{C}^{nn}$ (the set of $n \times n$ complex matrices) then the *graph* $G(A)$ of A is defined to be $(\langle n \rangle, E)$ where $\langle n \rangle = \{1, \cdots, n\}$ and $(i, j) \in E$ whenever $a_{ij} \neq 0$.

DEFINITION 3.4. Let $A \in \mathbb{C}^{nn}$ and let $\alpha = (i_0, e_1, i_1, \cdots, e_q, i_q)$ be a chain in $G(A)$. Then we define the *chain product* $\prod_\alpha(A)$ by

$$\prod_\alpha(A) = \prod_{p=1}^{q} a_{i_{p-1} i_p}^{e_p}.$$

We put $\prod_\varnothing(A) = 1$. If α is a cycle (path, circuit) we call the $\prod_\alpha(A)$ a cycle (path, circuit) product.

Note that if $\alpha = (i_0, e_1, i_1, \cdots, e_q, i_q)$ is a closed path and

$$\beta = (i_k, e_{k+1}, \cdots, e_q, i_0, e_1, \cdots, i_k), \qquad 0 \leqq k < q,$$

then $\prod_\alpha(A) = \prod_\beta(A)$.

DEFINITION 3.5. Let $A, B \in \mathbb{C}^{nn}$. We say that A and B are *diagonally similar* if there exists a nonsingular diagonal matrix X such that $B = X^{-1}AX$, and we say that A and B are *sign similar* if there exists a nonsingular diagonal sign matrix X such that $B = X^{-1}AX$. We say that A and B are *permutation similar* if there exists a permutation matrix P such that $B = P^{-1}AP$. We say that A and B are *diagonally equivalent* if there exist nonsingular diagonal matrices X and Y such that $B = YAX$.

DEFINITION 3.6. Let $A, B \in \mathbb{C}^{nn}$. We say that A and B are *c-equivalent* if $G(A) = G(B)$ and for all circuits α in $G(A)$ we have $\prod_\alpha(A) = \prod_\alpha(B)$.

Definition 3.6 and some implications may be found in [2]. In particular, it is well known that for irreducible matrices A and B, the matrices A and B are diagonally similar if and only if they are c-equivalent (see [2, Thm. 4.1]).

DEFINITION 3.7. If $V, W \subseteq \langle n \rangle$ and $A \in \mathbb{C}^{nn}$, then $A[V, W]$ is the submatrix of A whose rows are indexed by V and whose columns are indexed by W (in their natural orders).

DEFINITION 3.8. Let $A \in \mathbb{C}^{nn}$.

(i) The matrix A is called *irreducible* if $G(A)$ is strongly connected.

(ii) The matrix A is said to be in *Frobenius normal form* if A may be written in the block form

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1s} \\ 0 & A_{22} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & A_{ss} \end{bmatrix}, \tag{3.9}$$

where A_{ii} is an irreducible square matrix, $i = 1, \cdots, s$.

(iii) Let $B \in \mathbb{C}^{nn}$. The matrix B is said to be a *Frobenius normal form* of A if B is in Frobenius normal form and if A and B are permutation similar.

Remark 3.10. Let $A \in \mathbb{C}^{nn}$. We may obtain a Frobenius normal form of A by reordering the vertices of $G(A)$ so that V_p consists of consecutive integers, $p = 1, \cdots, s$, and so that (2.8) holds. It follows from Definition 3.8 that a Frobenius normal form of A is unique up to permutation similarity. The diagonal blocks of a Frobenius normal form of A will be called the *components* of A.

In [4, § 4] we introduced the inflation product $C \times\times A$ of two matrices where $C \in \mathbb{C}^{ss}$, $A \in \mathbb{C}^{nn}$, and A is partitioned into s blocks. In this paper we use the notation $C \times\times A$ only in the special case when A is in Frobenius normal form and C satisfies (3.12) below.

DEFINITION 3.11. Let $A \in \mathbb{C}^{nn}$ be in Frobenius normal form (3.9) and suppose that $C \in \mathbb{C}^{ss}$ satisfies

$$C \text{ is a sign matrix,} \tag{3.12}$$

$$c_{pp} \text{ is equal to 0 or 1,} \quad p \in \langle s \rangle, \tag{3.13}$$

$$c_{pq} = 0 \Leftrightarrow A_{pq} = 0, \quad p, q \in \langle s \rangle. \tag{3.14}$$

Then the matrix $B = C \times\times A \in \mathbb{C}^{nn}$ is defined to be the matrix with blocks $B_{pq} = c_{pq}A_{pq}$, $p, q \in \langle s \rangle$.

4. Preliminaries on equality classes and sufficient sets.

Notation 4.1. We use the following notation:

N = the set $\{0, 1, 2, \cdots\}$

Δ = the set $\{(i, i) : i \in \langle n \rangle\}$.

Notation 4.2. Let $A \in \mathbb{C}^{nn}$.

$\rho(A)$ = the spectral radius of A.

$\mathscr{U}_n$ = the set $\{A \in \mathbb{C}^{nn} : \rho(|A|) < 1\}$.

We normally write $\mathscr{U}$ in place of $\mathscr{U}_n$.

$\mathscr{J}$ = the set of irreducible matrices contained in $\mathscr{U}$.

If $G = (\langle n \rangle, E)$ is a graph, then $\mathscr{U}(G)$ is the set $\{A \in \mathscr{U} : G(A) = G\}$.

Note that for every $A \in \mathbb{C}^{nn}$ we have $cA \in \mathscr{U}$ for all complex numbers c whose absolute value is sufficiently small. Let $A \in \mathscr{U}$ and let $S \subseteq N$. Observe that

$$\left| \sum_{s \in S} A^s \right| \leqq \sum_{s \in S} |A|^s \leqq \sum_{s \in N} |A|^s = (I - |A|)^{-1}. \tag{4.3}$$

Hence the series in (4.3) converge. In order to discuss the cases when the equalities hold in (4.3) we shall make several definitions. The first of these allows us to discuss the case of equality in the first inequality in (4.3).

DEFINITION 4.4. Let $\Gamma \subseteq \langle n \rangle \times \langle n \rangle$, let $S \subseteq N$, and let $\mathscr{A} \subseteq \mathscr{U}$. Then the $(\mathscr{A}, \Gamma, S)$-*equality class* is defined to consist of all $A \in \mathscr{A}$ such that

$$\left(\left| \sum_{s \in S} A^s \right| \right)_{ij} = \left(\sum_{s \in S} |A|^s \right)_{ij}, \tag{4.5}$$

for all $(i, j) \in \Gamma$, and it is denoted by Equ $(\mathscr{A}, \Gamma, S)$.

The first two parameters in Equ $(\mathscr{A}, \Gamma, S)$ are optional and default to $\mathscr{U}$ and $\langle n \rangle \times \langle n \rangle$, respectively. Thus (by convention)

Equ (Γ, S) = Equ $(\mathscr{U}, \Gamma, S)$,

Equ $(\mathscr{A}, S)$ = Equ $(\mathscr{A}, \langle n \rangle \times \langle n \rangle, S)$,

Equ (S) = Equ $(\mathscr{U}, \langle n \rangle \times \langle n \rangle, S)$.

We have the following easy but fundamental lemma.

LEMMA 4.6. *Let $i, j \in \langle n \rangle$ and let $S \subseteq N$. Then the following conditions are equivalent*:

(i) $A \in \text{Equ}((i,j), S)$.

(ii) $\text{sgn}(\prod_\alpha(A)) = \text{sgn}(\prod_\beta(A))$, *for all paths* α, β *from* i *to* j *in* $G(A)$ *such that* $|\alpha|, |\beta| \in S$.

Proof. Note that (i) is equivalent to (4.5) by definition of Equ $((i,j), S)$. The equivalence of (i) and (ii) follows from the conditions for equality in the triangle inequality and the result that for $i, j \in \langle n \rangle$ and $s \in N$ we have

$$A_{ij}^s = \sum_{\sigma \in P(i,j;s)} \prod_\sigma(A) \tag{4.7}$$

where $P(i, j; s)$ is the set of all paths from i to j of length s in $G(A)$. □

The proof of the following lemma is easy and will be omitted.

LEMMA 4.8. *Let* $\mathscr{A} \subseteq \mathscr{B} \subseteq \mathscr{U}$ *and let* Γ, Γ', Λ *be subsets of* $\langle n \rangle \times \langle n \rangle$ *such that* $\Gamma \subseteq \Lambda$. *Let* $S \subseteq T \subseteq N$. *Then*

$$\text{Equ}(\mathscr{A}, \Gamma, S) = \text{Equ}(\mathscr{B}, \Gamma, S) \cap \mathscr{A}, \tag{4.9}$$

$$\text{Equ}(\mathscr{A}, \Gamma \cup \Gamma', S) = \text{Equ}(\mathscr{A}, \Gamma, S) \cap \text{Equ}(\mathscr{A}, \Gamma', S), \tag{4.10}$$

$$\text{Equ}(\mathscr{A}, \Lambda, T) \subseteq \text{Equ}(\mathscr{B}, \Gamma, S). \tag{4.11}$$

Let $S \subseteq N$. Then by Lemma 4.8 it follows that Equ $(\mathscr{A}, \Gamma, N) \subseteq$ Equ $(\mathscr{A}, \Gamma, S)$ for all $\mathscr{A} \subseteq \mathscr{U}$ and $\Gamma \subseteq \langle n \rangle \times \langle n \rangle$. This remark motivates the following definition which allows us to discuss the case of equality in the second inequality in (4.3).

DEFINITION 4.12. We say that the subset S of N is $(\mathscr{A}, \Gamma)$-*sufficient* if Equ $(\mathscr{A}, \Gamma, S)$ = Equ $(\mathscr{A}, \Gamma, N)$. We say that S is *minimal* $(\mathscr{A}, \Gamma)$-*sufficient* if S is $(\mathscr{A}, \Gamma)$-sufficient but no proper subset of S is $(\mathscr{A}, \Gamma)$-sufficient. We say that S is *optimal* $(\mathscr{A}, \Gamma)$-*sufficient* if S is an $(\mathscr{A}, \Gamma)$-sufficient of minimal cardinality, viz. there exists no $(\mathscr{A}, \Gamma)$-sufficient set of lower cardinality. The two parameters in the term (minimal, optimal) $(\mathscr{A}, \Gamma)$-sufficient are optional and default to $\mathscr{U}$ and $\langle n \rangle \times \langle n \rangle$, respectively. Thus S is Γ-sufficient means that S is $(\mathscr{U}, \Gamma)$-sufficient, S is $\mathscr{A}$-sufficient means that S is $(\mathscr{A}, \langle n \rangle \times \langle n \rangle)$-sufficient, S is sufficient means that S is $(\mathscr{U}, \langle n \rangle \times \langle n \rangle)$-sufficient.

Of course, an optimal $(\mathscr{A}, \Gamma)$-sufficient set is minimal $(\mathscr{A}, \Gamma)$-sufficient.

LEMMA 4.13. *Let* $\mathscr{A} \subseteq \mathscr{B} \subseteq \mathscr{U}$, *let* $\Gamma \subseteq \langle n \rangle \times \langle n \rangle$, *and let* $S \subseteq T \subseteq N$. *If* S *is* $(\mathscr{B}, \Gamma)$-*sufficient then* T *is* $(\mathscr{A}, \Gamma)$-*sufficient.*

Proof. By Lemma 4.8 we have

$$\text{Equ}(\mathscr{B}, \Gamma, N) \subseteq \text{Equ}(\mathscr{B}, \Gamma, T) \subseteq \text{Equ}(\mathscr{B}, \Gamma, S).$$

But by our hypothesis Equ $(\mathscr{B}, \Gamma, S)$ = Equ $(\mathscr{B}, \Gamma, N)$ and it follows that

$$\text{Equ}(\mathscr{B}, \Gamma, T) = \text{Equ}(\mathscr{B}, \Gamma, N).$$

Therefore, by (4.9), it follows that

$$\text{Equ}(\mathscr{A}, \Gamma, T) = \text{Equ}(\mathscr{B}, \Gamma, T) \cap \mathscr{A} = \text{Equ}(\mathscr{B}, \Gamma, N) \cap \mathscr{A} = \text{Equ}(\mathscr{A}, \Gamma, N). \quad \square$$

5. The equality class of *N*. In this section we prove necessary and sufficient conditions for $A \in$ Equ (Γ, N) for irreducible and general $A \in \mathscr{U}$. In view of Definition 4.4, $A \in$ Equ (Γ, N) is equivalent to

$$|(I - A)^{-1}|_{ij} = (I - |A|)_{ij}^{-1} \quad \text{for } (i,j) \in \Gamma. \tag{5.1}$$

THEOREM 5.2. *Let* $i, j \in \langle n \rangle$, *and let* Λ *be a subset of* $\langle n \rangle \times \langle n \rangle$ *such that* $(i, j) \in \Lambda$ *and* (i, j) *access covers* Λ. *Let* $A \in \mathscr{U}$. *Then the following conditions are equivalent.*

(i) $A \in \text{Equ}((i,j), N)$.

(ii) $\text{sgn}(\prod_\alpha(A)) = \text{sgn}(\prod_\beta(A))$, *for all paths* α, β *from* i *to* j *in* $G(A)$.

(iii) $\operatorname{sgn}(\prod_\alpha(A)) = \operatorname{sgn}(\prod_\beta(A))$, *for all paths* α, β *from* h *to* k *in* $G(A)$, *where* $(h, k) \in \Lambda$.

(iv) $A \in \operatorname{Equ}(\Lambda, N)$.

(v) *Both*

(a) $\operatorname{sgn}(\prod_\beta(A)) = \operatorname{sgn}(\prod_\gamma(A))$, *for all simple paths* β, γ *from* i *to* j *in* $G(A)$ *and*

(b) *If* α *is a circuit of* $G(A)$ *which is* $G(A)$*-access covered by* (i, j) *then* $\prod_\alpha(A) > 0$.

(vi) *All chain products of two-twisted closed chains of* $G(A)$ *which are* $G(A)$*-access covered by* (i, j) *are positive.*

(vii) *All cycle products of two-twisted cycles of* $G(A)$ *which are* $G(A)$*-access covered by* (i, j) *are positive.*

Proof. We shall show that (i) $\Leftrightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (iv) $\Rightarrow$ (i), (ii) $\Leftrightarrow$ (v), and (ii) $\Leftrightarrow$ (vi) $\Leftrightarrow$ (vii).

(i) $\Leftrightarrow$ (ii). This is given by Lemma 4.6.

(ii) $\Rightarrow$ (iii). Suppose (ii) holds. Let $(h, k) \in \Lambda$ and let α and β be paths from h to k in $G(A)$. Since (i, j) is a $G(A)$-access cover for (h, k) there exist paths γ and δ in $G(A)$ from i to h and k to j, respectively. Since

$$\prod_{\gamma\alpha\delta}(A) = \prod_{\gamma\beta\delta}(A)$$

by (ii), and since

$$\prod_{\gamma\alpha\delta}(A) = \prod_\gamma(A) \prod_\alpha(A) \prod_\delta(A),$$

$$\prod_{\gamma\beta\delta}(A) = \prod_\gamma(A) \prod_\beta(A) \prod_\delta(A),$$

we obtain (iii).

(iii) $\Rightarrow$ (iv). By (4.10) and Lemma 4.6.

(iv) $\Rightarrow$ (i). By (4.11), since $(i, j) \in \Lambda$.

(ii) $\Rightarrow$ (v). Suppose (ii) holds. Then obviously we have (a). To prove (b), let $\alpha = (i_0, \cdots, i_s)$ be a circuit of $G(A)$ that is $G(A)$-access covered by (i, j). Then there is a vertex k of α for which there exist paths δ from i to k and ψ from k to j. Without loss of generality we may assume that $k = i_0$. By (ii) the path products corresponding to the paths $\delta\psi$ and $\delta\alpha\psi$ have the same (nonzero) sign. It follows that $\prod_\alpha(A) > 0$ and (v) is proved.

(v) $\Rightarrow$ (ii). Suppose that (a) and (b) hold. Let δ be a path in $G(A)$ from i to j. Then $\prod_\delta(A)$ is a product of $\prod_\beta(A)$ and factors of type $\prod_\alpha(A)$, where β is a simple path from i to j and α is a circuit of $G(A)$ for which (i, j) is a $G(A)$-access cover. By (b), $\operatorname{sgn}(\prod_\delta(A)) = \operatorname{sgn}(\prod_\beta(A))$. Hence it follows from (a) that products corresponding to every pair of paths from i to j have the same sign.

(ii) $\Rightarrow$ (vi). Let $\alpha = (i_0, e_1, \cdots, i_m)$ with $i_0 = i_m$ be a two-twisted closed chain which is $G(A)$-access covered by (i, j). If $t(\alpha) = 0$ then the positivity of $\prod_\alpha(A)$ follows as in the proof of (ii) implies (v) with "circuit" replaced by "closed path." Suppose $t(\alpha) = 2$. Let α have twists at p and q, respectively. Without loss of generality we may assume that $p = 0$ and $e_1 = 1$. Observe that $e_{q+1} = -1$. Let $\alpha_1 = (i_0, \cdots, i_q)$ and let $\alpha_2 = (i_q, \cdots, i_s)^*$. Observe that both α_1 and a_2 are paths from i_0 to i_q. Since (i, j) is a $G(A)$-access cover for α, there exists paths δ from i to i_0 and ψ from i_q to j. By (ii), the nonzero path products corresponding to $\delta\alpha_1\psi$ and $\delta\alpha_2\psi$ have the same sign. Thus the path products corresponding to α_1 and α_2 have the same sign. Since $\alpha = \alpha_1\alpha_2^*$ our claim follows.

(vi) $\Rightarrow$ (ii). Let α and β be two paths from i to j in $G(A)$. Then $\alpha\beta^*$ is a two-twisted closed chain (possibly trivial). Since

$$\prod_{\alpha\beta^*}(A) = \prod_{\alpha}(A)/\prod_{\beta}(A)$$

clearly (vi) implies (ii).

(vi) $\Rightarrow$ (vii). This is trivial.

(vii) $\Rightarrow$ (vi). Assume that (vii) holds and let $\alpha = (i_0, \cdots, i_s)$, $i_0 = i_s$, be a two-twisted closed chain which is $G(A)$-access covered by (i, j). The proof is by induction on the length s. If $s = 1$, then α is a cycle and the result holds. So let $s > 1$ and assume that $\prod_\gamma(A) > 0$ for every two-twisted closed chain γ that is $G(A)$-access covered by (i, j) and such that $|\gamma| < s$. If α is a cycle the result holds. Otherwise, there exist p and q, $0 \leqq p < q < s$, such that $\delta = (i_p, \cdots, i_q)$ is a cycle. Further, $\beta = (i_0, \cdots, i_p, i_{q+1}, \cdots, i_s)$ is a closed chain of length less than s which is $G(A)$-access covered by (i, j). By Lemma 2.12, δ and β are two-twisted and hence by the inductive assumption the corresponding chain products are positive. But $\prod_\alpha(A) = \prod_\beta(A) \prod_\delta(A)$, and hence $\prod_\alpha(A) > 0$. We now deduce (vi). □

It is easy to construct an example to show that the assumption $(i, j) \in \Lambda$ cannot be omitted from the hypothesis of Theorem 5.2. However, we have the following corollary.

COROLLARY 5.3. *Let $A \in \mathbb{C}^{nn}$ and let $i, j, h, k \in \langle n \rangle$. Let (i, j) be a $G(A)$-access cover for (h, k). Then* Equ $((i, j), N) \subseteq$ Equ $((h, k), N)$.

Proof. Let $A \in$ Equ $((i, j), N)$. Let α and β be paths in $G(A)$ from h to k. Since (i, j) $G(A)$-access covers (h, k), there exist paths γ from i to h and δ from k to j in $G(A)$. By Theorem 5.2,

$$\prod_{\gamma\alpha\delta}(A) = \prod_{\gamma\beta\delta}(A)$$

and it follows that

$$\prod_{\alpha}(A) = \prod_{\beta}(A).$$

Hence, by Theorem 5.2, $A \in$ Equ $((h, k), N)$. □

COROLLARY 5.4. *Let $G = (\langle n \rangle, E)$ be a graph and let $\Gamma \subseteq \Lambda \subseteq A_G(\Gamma) \subseteq \langle n \rangle \times \langle n \rangle$. Then*

$$\text{Equ } (\mathcal{U}(G), \Lambda, N) = \text{Equ } (\mathcal{U}(G), \Gamma, N).$$

Proof. Since $\Gamma \subseteq \Lambda$, it follows from (4.11) that

$$\text{Equ } (\mathcal{U}(G), \Lambda, N) \subseteq \text{Equ } (\mathcal{U}(G), \Gamma, N).$$

Hence we need only prove that

$$\text{Equ } (\mathcal{U}(G), \Gamma, N) \subseteq \text{Equ } (\mathcal{U}(G), \Lambda, N). \tag{5.5}$$

By (4.10) we have

$$\text{Equ } (\mathcal{U}(G), \Gamma, N) = \cap \{\text{Equ } (\mathcal{U}(G), (i,j), N) : (i,j) \in \Gamma\}. \tag{5.6}$$

and similarly

$$\text{Equ } (\mathcal{U}(G), \Lambda, N) = \cap \{\text{Equ } (\mathcal{U}(G), (h,k), N) : (h,k) \in \Lambda\}. \tag{5.7}$$

It follows from the definition of $A_G(\Gamma)$ that for each $(h, k) \in \Lambda$ there exists $(i, j) \in \Gamma$ such that (i, j) G-access covers (h, k). Hence (5.5) now follows from (5.6), (5.7), and Corollary 5.3. □

As a special case of Theorem 5.2 we obtain the following corollary, which is essentially known.

COROLLARY 5.8. *Let $A \in \mathcal{U}$. Then the following are equivalent*:

(i) $A \in \text{Equ}(\Delta, N)$.

(ii) *Every circuit product for $G(A)$ is positive.*

Proof. (i) $\Rightarrow$ (ii). Let $A \in \text{Equ}(\Delta, N)$. Since Δ is a $G(A)$-access cover for every circuit it follows by Theorem 5.2, Part (v) that every circuit product is positive.

(ii) $\Rightarrow$ (i). This follows from Theorem 5.2, Part (v), since the only simple paths from i to i, $i \in \langle n \rangle$, are circuits. □

For irreducible matrices there is the following stronger result which is essentially due to Neumaier [6] and which motivated our investigations.

COROLLARY 5.9. *Let Γ be a nonempty subset of $\langle n \rangle \times \langle n \rangle$ and let $A \in \mathcal{J}$. Then the following are equivalent*:

(i) $A \in \text{Equ}(N)$.

(ii) $A \in \text{Equ}(\Gamma, N)$.

(iii) *All circuit products of $G(A)$ are positive.*

(iv) *All closed path products of $G(A)$ are positive.*

(v) *A is sign similar to $|A|$.*

Proof. (i) $\Rightarrow$ (ii). This implication follows from Lemma 4.8.

(ii) $\Rightarrow$ (iii). Suppose that (ii) holds. Since $G(A)$ is strongly connected, it follows from Lemma 2.7 that Γ is a $G(A)$-access cover for $\langle n \rangle$ and (iii) follows immediately from Theorem 5.2.

(iii) $\Rightarrow$ (iv). Every closed path product is a product of circuit products.

(iv) $\Rightarrow$ (v). Suppose (iv) holds. Then corresponding circuit products of A and $|A|$ are equal. Thus, since A is irreducible, as is well known (e.g., [2, Thm. 4.1]), there exists a diagonal matrix X such that $X^{-1}AX = |A|$. Let $D = |X^{-1}|X$. Then D is a diagonal sign matrix satisfying $D^{-1}AD = |A|$.

(v) $\Rightarrow$ (i). Let D be a diagonal sign matrix such that $D^{-1}AD = |A|$. Since $|A|^k = D^{-1}A^kD$ and $\rho(A) < 1$, it follows that

$$D^{-1}(I-A)^{-1}D = (I-|A|)^{-1}.$$

Hence, since D is a diagonal sign matrix, (5.1) holds for $\Gamma = \langle n \rangle \times \langle n \rangle$ and (i) is proved. □

LEMMA 5.10. *Let $A \geqq 0$ be an $n \times n$ matrix in Frobenius normal form and let C be an (upper triangular) $s \times s$ matrix satisfying* (3.12)–(3.14). *Let $B = C \times\times A$. Let $i, j \in \langle n \rangle$ and suppose that a_{ij} is an element of A_{pq}, where $1 \leqq p, q \leqq s$. Then for every path β in $G(B)$ from i to j there is a path γ in $G(C)$ from p to q such that*

$$\text{sgn}\,(\prod_\beta(B)) = \prod_\gamma(C). \tag{5.11}$$

Conversely, for every path γ in $G(C)$ from p to q there is a path β in $G(B)$ from i to j such that (5.11) *is satisfied.*

Proof. Suppose the rows and columns of the component A_{rr} of A are indexed by the subset V_r of $\langle n \rangle$, $r = 1, \cdots, s$. Since A and B are in Frobenius normal form, there exist p_t, $t = 0, \cdots, k$, $1 \leqq p_t \leqq s$ with $p_0 = p$ and $p_k = q$ and i_t, j_t in V_{p_t}, $t = 0, \cdots, k$, with $i_0 = i$ and $j_k = j$, such that

$$\beta = \beta_0\delta_1\beta_1 \cdots \beta_k, \tag{5.12}$$

where β_t is a path from i_t to j_t in $G(B_{p_tp_t})$, $t = 0, \cdots, k$ and $\delta_t = j_{t-1} \to i_t$, $t = 1, \cdots, k$. Since $A \geqq 0$ and $c_{p_tp_t} = 1$ or 0, $t = 1, \cdots, k$ and $c_{p_tp_t} = 0$ if and only if $B_{p_tp_t}$ is a zero 1×1 block in which case β_t is empty, we have $\prod_{\beta_t}(B) > 0$ and $\prod_{\delta_t}(B) = c_{p_tp_t}$. Hence if we define

$$\gamma = p_0 \to \cdots \to p_k, \tag{5.13}$$

then γ is a path in $G(C)$ from p to q such that (5.11) holds.

Conversely, let $B = C \times\times |A|$ and let γ given by (5.13) be a path in $G(C)$ from p to q. We may choose $i_t, j_t \in V_{p_t}$, $t = 0, \cdots, k$ and paths β_t from i_t to j_t in $G(B_{p_t})$, $t = 0, \cdots, k$. If δ_t is again defined to be $j_{t-1} \to i_t$, $t = 1, \cdots, k$ and β is defined by (5.12) then (5.11) holds, since $c_{pp} \geqq 0$, $p = 1, \cdots, s$. □

We now apply Theorem 5.2 to obtain the final result in this section.

THEOREM 5.14. *Let $A \in \mathcal{U}$ and let Γ be a $G(A)$-access cover. Then the following are equivalent.*

(i) $A \in \text{Equ}(\Gamma, N)$.

(ii) $\text{sgn}(\prod_\alpha(A)) = \text{sgn}(\prod_\beta(A))$, *for all paths α, β in $G(A)$ from i to j, where $(i, j) \in \Gamma$.*

(iii) $\text{sgn}(\prod_\alpha(A)) = \text{sgn}(\prod_\beta(A))$, *for all paths α, β in $G(A)$ from i to j, where $(i, j) \in \langle n \rangle \times \langle n \rangle$.*

(iv) $A \in \text{Equ}(N)$.

(v) *Both*

(a) $\text{sgn}(\prod_\beta(A)) = \text{sgn}(\prod_\gamma(A))$, *for all simple paths β, γ in $G(A)$ from i to j, where $(i, j) \in \Gamma$,*

and

(b) $\prod_\alpha(A) > 0$ *for all circuits α of $G(A)$.*

(v′) *Both*

(a′) $\text{sgn}(\prod_\beta(A)) = \text{sgn}(\prod_\gamma(A))$, *for all simple paths β, γ in $G(A)$ from i to j, where $(i, j) \in \langle n \rangle \times \langle n \rangle$,*

and

(b′) $\prod_\alpha(A) > 0$ *for all circuits α of $G(A)$.*

(vi) *All chain products of two-twisted closed chains of $G(A)$ are positive.*

(vii) *All cycle products of two-twisted cycles are positive.*

(viii) *If A is in Frobenius normal form* (3.9) *then there exists an $s \times s$ sign matrix C such that $zC \in \text{Equ}(\mathcal{U}_s, N)$, for $0 < z < 1$ and A is sign similar to $C \times\times |A|$.*

Proof. The equivalence of conditions (i)–(vii) follows immediately from the equivalence of the correspondingly numbered conditions in Theorem 5.2 and the fact that

$$\text{Equ}(\Gamma, N) = \cap \{\text{Equ}((i,j), N) : (i,j) \in \Gamma\}$$

by (4.10). The equivalence of conditions (v) and (v′) is easily derived by means of Conditions (iv) and (v) of Theorem 5.2. So it suffices to prove the equivalence of Conditions (iv) and (viii).

(iv) $\Rightarrow$ (viii). Suppose that (iv) holds. Since A_{pp} is irreducible, $p = 1, \cdots, s$, by Corollary 5.9 there exist diagonal sign matrices X_p that satisfy $X_p^{-1} A X_p = |A_{pp}|$, $p = 1, \cdots, s$. Let $X = X_1 \oplus \cdots \oplus X_s$ and let $B = X^{-1} A X$. Then $|B| = |A|$ and $B_{pp} \geqq 0$, $p = 1, \cdots, s$. We shall show that $B = C \times\times |A|$, where C is a suitably chosen sign matrix satisfying conditions (3.12)–(3.14).

Let $1 \leqq i, j, h, k \leqq n$ and suppose that both b_{ij} and b_{hk} are nonzero elements of B_{pq}, where $1 \leqq p, q \leqq s$. Since B_{pp} and B_{qq} are irreducible, there exist chains α and γ in $G(B_{pp})$ from i to h and in $G(B_{qq})$ from k to j, respectively. Since $B_{pp} \geqq 0$ and $B_{qq} \geqq 0$, the products $\prod_\alpha(B)$ and $\prod_\gamma(B)$ are positive. Let β, δ be chains $h \to k$ and $i \to j$ of length 1, respectively. Then $\alpha\beta\gamma$ and δ are paths from i to j in $G(B)$. Since $A \in \text{Equ}(N)$ we also have $B \in \text{Equ}(N)$ and it follows from (ii) of Theorem 5.2 that

$$\text{sgn}(\prod_{\alpha\beta\gamma}(B)) = \text{sgn}(\prod_\delta(B)).$$

We deduce that $\text{sgn}(b_{hk}) = \text{sgn}(b_{ij})$.

Thus we may define

$$c_{pq} = \begin{cases} 0 & \text{if } B_{pq} = 0 \\ \operatorname{sgn}(b_{ij}) & \text{if } B_{pq} \neq 0, \text{ where } b_{ij} \text{ is any nonzero entry of } B_{pq}. \end{cases}$$

Then c_{pp} is equal to 0 or 1 since $B_{pp} \geqq 0$, $p = 1, \cdots, s$. Thus $C \in \mathbb{C}^{ss}$ is an (upper triangular) matrix that satisfies conditions (3.12)–(3.14). Further, $B = C \times\times |A|$.

We must still show that $zC \in \operatorname{Equ}(\mathcal{U}, N)$ for $0 < z < 1$. Let $p, q \in \langle s \rangle$ and let γ be a chain from p to q in $G(C)$. Let i and j be elements of the sets V_p and V_q (which index the corresponding components), respectively. Then by Lemma 5.10 there exists a chain from i to j in $G(B)$ such that (5.11) holds. It follows that path products corresponding to any two paths from p to q in $G(B)$ have the same sign. Let $0 < z < 1$. Since $\rho(zC) < 1$, we now obtain $zC \in \operatorname{Equ}(N)$ by Theorem 5.2.

(viii) $\Rightarrow$ (iv). Suppose that (viii) holds and put $B = C \times\times |A|$. Let $i, j \in \langle n \rangle$ and let α, β be paths from i to j in $G(B)$. It follows from Lemma 5.10 that there exists a path γ in $G(C)$ such that

$$\operatorname{sgn}(\textstyle\prod_\alpha(B)) = \operatorname{sgn}(\prod_\gamma(C)) = \operatorname{sgn}(\prod_\beta(B)).$$

Hence $B \in \operatorname{Equ}(N)$ by Theorem 5.2. Since A is sign similar to B we obtain (iv). □

For the terminology and definitions employed in the following remark see [5].

Remark 5.15. (i) Our proof of (vii) $\Rightarrow$ (vi) of Theorem 5.2 shows that every algebraic two-twisted chain in a graph G is an integral linear combination of algebraic two-twisted cycles.

(ii) Suppose that $A \in \mathcal{U}$ and let W be the subspace of the flow space of $G(A)$ which is spanned by the algebraic two-twisted closed chains of $G(A)$. Let X be an integral spanning set for W. If the chain products corresponding to the closed chains in X are positive, then all chain products corresponding to chains in W are positive. Hence (vi) of Theorem 5.14 holds, and it follows that $A \in \operatorname{Equ}(N)$. However, this conclusion does not follow for arbitrary (nonintegral) spanning sets as one may see from Example 5.2 in [8]. A similar remark may be made concerning (vi) of Theorem 5.2.

6. Sufficient sets. We begin this section with some applications of Corollary 5.4.

COROLLARY 6.1. *Let $G = (\langle n \rangle, E)$ be a graph. Let $\Gamma \subseteq \Lambda \subseteq \langle n \rangle \times \langle n \rangle$ and suppose that Γ is a G-access cover for Λ. Let $S \subseteq N$. If S is $(\mathcal{U}(G), \Gamma)$-sufficient then S is $(\mathcal{U}(G), \Lambda)$-sufficient.*

Proof. By (4.11) we have

$$\operatorname{Equ}(\mathcal{U}(G), \Lambda, S) \subseteq \operatorname{Equ}(\mathcal{U}(G), \Gamma, S). \tag{6.2}$$

By assumption,

$$\operatorname{Equ}(\mathcal{U}(G), \Gamma, S) = \operatorname{Equ}(\mathcal{U}(G), \Gamma, N), \tag{6.3}$$

and by Corollary 5.4,

$$\operatorname{Equ}(\mathcal{U}(G), \Gamma, N) = \operatorname{Equ}(\mathcal{U}(G), \Lambda, N). \tag{6.4}$$

It follows from (6.2)–(6.4) that

$$\operatorname{Equ}(\mathcal{U}(G), \Lambda, S) \subseteq \operatorname{Equ}(\mathcal{U}(\Gamma), \Lambda, N).$$

But hence by (4.11) we obtain

$$\operatorname{Equ}(\mathcal{U}(G), \Lambda, S) = \operatorname{Equ}(\mathcal{U}(\Gamma), \Lambda, N)$$

which proves the corollary. □

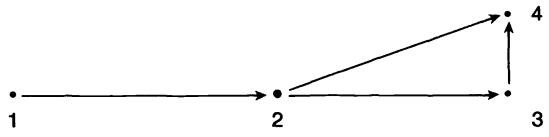


FIG. 1

Example 6.5. Let $\langle n\rangle = 4$ and let G be given by Fig. 1. Let $S_1 = \{2, 3\}$. Then S_1 is $(\mathscr{U}(G), (1, 4))$-sufficient but not $(\mathscr{U}(G), (2, 4))$-sufficient. Since $(1, 4)$ is a G-access cover for $(2, 4)$, this shows that the condition $\Gamma \subseteq \Lambda$ cannot be omitted in Corollary 6.1.

Next let $S_2 = \{0\}$ and let $\Gamma = \{(3, 4)\}$. Then S_2 is $(\mathscr{U}(G), \Gamma)$-sufficient but not Λ-sufficient if $(2, 4) \in \Lambda \subseteq \langle n\rangle \times \langle n\rangle$. By choosing $\{(2, 4), (3, 4)\} \subseteq \Lambda$, we obtain an example with S_2 is $(\mathscr{U}(G), \Gamma)$-sufficient but not $(\mathscr{U}(G), \Lambda)$-sufficient even though $\Gamma \subseteq \Lambda$, and thus the condition that Γ is a G-access cover for Λ cannot be omitted in Corollary 6.1. Finally, observe that S_1 is $(\mathscr{U}(G), \Lambda)$-sufficient for any set Λ such that $(1, 4) \in \Lambda \subseteq \langle n\rangle \times \langle n\rangle$. Choosing $\{(1, 4), (2, 4)\} \subseteq \Lambda$ and putting $\Gamma = \{(2, 4)\}$ it follows from our previous remarks that S_1 is $(\mathscr{U}(G), \Lambda)$-sufficient, but not $(\mathscr{U}(G), \Gamma)$-sufficient. Note that $\Gamma \subseteq \Lambda$. Thus there appears to be no simple relation in general (without the condition that Γ is a G-access cover for Λ) between $(\mathscr{U}(G), \Gamma)$-sufficiency and $(\mathscr{U}(G), \Lambda)$-sufficiency when $\Gamma \subseteq \Lambda$.

We shall give two proofs of our next corollary. The first is an application of Corollary 6.1 and the second is based directly on Lemma 4.6.

COROLLARY 6.6. *Let* $\Gamma \subseteq \Lambda \subseteq \langle n\rangle \times \langle n\rangle$ *where* $\Gamma \neq \varnothing$. *Let* $S \subseteq N$. *If* S *is* $(\mathscr{J}, \Gamma)$-*sufficient then* S *is* $(\mathscr{J}, \Lambda)$-*sufficient.*

First proof. Let $A \in \text{Equ}(\mathscr{J}, \Lambda, S)$. Then $A \in \text{Equ}(\mathscr{U}(G(A)), \Lambda, S)$. Since

$$\mathscr{U}(G(A)) \subseteq \mathscr{J},$$

it follows from Lemma 4.13 that S is $(\mathscr{U}(G(A)), \Gamma)$-sufficient. But since $G(A)$ is strongly connected it follows from Lemma 2.7 that Γ is a $G(A)$-access cover for Λ. Hence, by Corollary 6.1, S is $(\mathscr{U}(G(A)), \Lambda)$-sufficient. It follows that

$$A \in \text{Equ}(\mathscr{U}(G(A)), \Lambda, N) \subseteq \text{Equ}(\mathscr{J}, \Lambda, N).$$

The result follows.

Second proof. Let $A \in \text{Equ}(\mathscr{J}, \Lambda, S)$. Then, by Lemma 4.6, for all $(i, j) \in \Gamma$, $\operatorname{sgn} \prod_\alpha(A) = \operatorname{sgn} \prod_\beta(A)$, for all paths α, β from i to j in $G(A)$ such that $|\alpha|, |\beta| \in S$. Hence, since S is $(\mathscr{J}, \Gamma)$-sufficient, it follows that $A \in \text{Equ}(\mathscr{J}, \Gamma, N)$ and consequently $\operatorname{sgn} \prod_\alpha(A) = \operatorname{sgn} \prod_\beta(A)$ for all paths α, β from i to j in $G(A)$, where $(i, j) \in \Gamma$, without restriction on the lengths of α and β. Hence, also, $\operatorname{sgn} \prod_\gamma(A) = \operatorname{sgn} \prod_\delta(A)$ for all paths γ, δ from h to k in $G(A)$, where $(h, k) \in \Lambda$, since, by Lemma 2.7, these paths can be extended to paths α, β, respectively, from i to j with $(i, j) \in \Gamma$. But this proves that S is $(\mathscr{J}, \Gamma)$-sufficient. □

Of course, the most interesting case of Corollary 6.6 arises when

$$\{(i,j)\} = \Gamma \subseteq \Lambda = \langle n\rangle \times \langle n\rangle.$$

DEFINITION 6.7. Let S be a nonempty subset of N. Then we define

$D(S) = \{s - t: s, t \in S \text{ and } s > t\}$,
$\gcd(S)$ = the greatest common divisor of the elements of S,
$C(S) = \{\gcd(T): T \subseteq S, T \neq \varnothing\}$,

$CD(S) = C(D(S))$,
$D(\varnothing) = C(\varnothing) = \varnothing$.

Observe that $S \subseteq C(S)$. For example, if $S = \{3, 9, 13, 18\}$ then

$$D(S) = \{4, 5, 6, 9, 10, 15\}$$

and $CD(S) = \{1, 2, 3, 4, 5, 6, 9, 10, 15\}$. Note also that $C(C(S)) = C(S)$. Since $C(CD(S)) = CD(S)$, it follows that every element of $CD(S)$ is a multiple of the minimal element of $CD(S)$.

LEMMA 6.8. *Let $S \subseteq N$ and let $A \in \text{Equ}(S)$. Then for a closed path α in $G(A)$ with length $s \in CD(S)$ we have $\prod_\alpha(A) > 0$.*

Proof. Let

$$\alpha = i_0 \to \cdots \to i_{s-1} \to i_0.$$

We first show $\prod_\alpha(A) > 0$ for $s \in D(S)$. Then $s = v - u$, where $u, v \in S$. Write $u = as + t$, where a and t are nonnegative integers and $t < s$. Then $v = (a + 1)s + t$. We take $\beta[\gamma]$ to be the path from i_0 to i_t of length $u[v]$ obtained by repeating $a[a + 1]$ times the path α and adjoining $i_0 \to \cdots \to i_t$. Since $A \in \text{Equ}(S)$, it follows from Lemma 4.6 that the nonzero path products $\prod_\beta(A)$ and $\prod_\gamma(A) = \prod_\beta(A)\prod_\alpha(A)$ have the same sign. Hence $\prod_\alpha(A) > 0$.

We now consider the general case of $s \in CD(S)$. Then there exist $s_1, s_2, \cdots, s_k$ in $D(S)$ whose gcd is s. As is well known, there exist integers a_i, $i = 1, \cdots, k$, such that

$$s = \sum_{i=1}^{k} a_i s_i. \tag{6.9}$$

Without loss of generality, assume that $a_i \leqq 0$ if and only if $1 \leqq i \leqq q$. Let ω_i be the closed path from i_0 to i_0 obtained by repeating s_1/s times the path α. By the first part of the proof, ω_i has a positive path product. Let $\mu[\nu]$ be the closed path from i_0 to i_0 obtained by repeating $|a_i|$ times the path ω_i, $i = 1, \cdots, t$ $[i = t + 1, \cdots, k]$. By (6.9), ν is obtained by adjoining α to μ. Since μ and ν have positive path products it follows that $\prod_\alpha(A) > 0$. □

COROLLARY 6.10. *If $A \in \text{Equ}(S)$ then $A \in \text{Equ}(\Delta, CD(S))$.*

Proof. Immediate by Lemmas 6.8 and 4.6. □

The converses of Lemma 6.8 and Corollary 6.10 are false if $n > 1$ even for irreducible matrices. In fact, we shall give an example of an irreducible matrix A and a set S, for which every closed path of length $s \in CD(S)$ has positive path product, yet the matrix A is not even in $\text{Equ}((i, i), S)$, for any $i \in \langle n \rangle$.

Example 6.11. Let A be the $n \times n$ matrix with all entries on and above the diagonal equal to 1 and all entries below the diagonal equal to -1. Let $S = \{1, 2\}$ and let $i, j \in \langle n \rangle$, $i \neq j$. Observe that the circuit product corresponding to $i \to i$ is positive while the circuit product corresponding to $i \to j \to i$ is negative. Hence, by Lemma 4.6, $A \notin \text{Equ}((i, i), S)$. However $CD(S) = \{1\}$ and all circuit products of length 1 are positive. Hence $A \in \text{Equ}(CD(S))$, by Lemma 4.6.

THEOREM 6.12. *Let S be a subset of N. Then the following are equivalent.*

(i) *S is Δ-sufficient.*
(ii) *S is $(\mathcal{J}, \Delta)$-sufficient.*
(iii) *S is $\mathcal{J}$-sufficient.*
(iv) *$CD(S)$ contains $\langle n \rangle$.*
(v) *For all $A \in \text{Equ}(S)$, all circuit products of A are positive.*
(vi) *For all $A \in \text{Equ}(S)$, A is diagonally similar to a matrix B such that all irreducible diagonal blocks in the Frobenius normal form of B are nonnegative.*

Proof. (i) $\Rightarrow$ (ii) is obvious.

(ii) $\Rightarrow$ (iii). By Corollary 6.6.

(iii) $\Rightarrow$ (iv). Let $k \in \langle n \rangle$ and let λ be a nonzero complex number. Suppose that $k \notin CD(S)$. We shall prove the claimed implication by constructing an irreducible $n \times n$ matrix $A(k, \lambda)$ such that, for suitable λ, $A(k, \lambda) \in \text{Equ}(\mathcal{J}, S) \backslash \text{Equ}(\mathcal{J}, N)$. If $k = 1$, we let $A(k, \lambda)$ be the $n \times n$ matrix all of whose entries are λ. If $k \in \{2, \cdots, n\}$ we define $A(k, \lambda) = A$ by

$$a_{i,i+1} = 1, \quad i = 1, \cdots, k-2,$$

$$a_{k-1,j} = \lambda, \quad j = k, \cdots, n,$$

$$a_{j,1} = 1, \quad j = k, \cdots, n,$$

$$a_{ij} = 0, \quad \text{otherwise}, \quad i, j \in \langle n \rangle,$$

e.g., for $n = 5$ and $k = 4$

$$A(k, \lambda) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \lambda & \lambda \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Note that for all $k \in \langle n \rangle$ the matrix $A(k, \lambda)$ is irreducible and the length of every closed path of $A(k, \lambda)$ is a multiple of k and, provided that $k \geqq 2$, every circuit product of $A(k, \lambda)$ equals λ. For all $k \in \langle n \rangle$, it follows that for every closed path δ of $G(A(k, \lambda))$ we have

$$\prod_\delta (A(k, \lambda)) = \lambda^h, \quad \text{where} \quad |\delta| = hk. \tag{6.13}$$

We now choose λ depending on two cases.

(a) No multiple of k lies in $CD(S)$. Then let $\lambda = -1$.

(b) Some positive multiple of k is in $CD(S)$. Then let pk be the smallest such multiple and λ be a primitive pth root of unity. Since $k \notin CD(S)$ we have $p > 1$.

Let $i, j \in \langle n \rangle$ and let α and β be paths from i to j in $G(A)$. Suppose that $|\alpha|$ and $|\beta|$ belong to S and assume without loss of generality that $|\alpha| \geqq |\beta|$. Let $d = |\alpha| - |\beta|$. Let γ be a path from j to i in $G(A(k, \lambda))$, which exists since $A(k, \lambda)$ is irreducible. Observe that $\alpha\gamma$ and $\beta\gamma$ are closed paths and hence $d = |\alpha\gamma| - |\beta\gamma|$ is divisible by k.

Suppose first that $d = 0$. Then $\alpha\gamma$ and $\beta\gamma$ are closed paths of the same length. It follows from (6.13) that the closed path products corresponding to $\alpha\gamma$ and $\beta\gamma$ are equal. Suppose now that $d > 0$. Then $d \in D(S) \subseteq CD(S)$. Hence (b) above holds. We recall that $C(CD(S)) = CD(S)$. Hence, since pk is the minimal multiple of k in $CD(S)$, it follows that d must be a multiple of pk. But (6.13) then again implies that the closed path products corresponding to $\alpha\gamma$ and $\beta\gamma$ are equal. Hence, in either case, $\prod_\alpha(A) = \prod_\beta(A)$. Since i, j are arbitrary in $\langle n \rangle$, it follows from Lemma 4.6 that $A \in \text{Equ}(S)$.

On the other hand, since $A(k, \lambda)$ has a circuit α of length k and $\prod_\alpha(A(k, \lambda))$ is not positive, we have by Theorem 5.2 that $A(k, \lambda) \notin \text{Equ}(N)$. The implication (iii) $\Rightarrow$ (iv) is proved.

(iv) $\Rightarrow$ (v). Immediate by Lemma 6.8.

(v) $\Rightarrow$ (vi). By Fiedler and Ptak [3] or Engel and Schneider [2] an irreducible matrix that satisfies (v) is diagonally similar to a nonnegative matrix. By applying this result to the Frobenius normal form of A we obtain (vi) from (v).

(vi) $\Rightarrow$ (i). Let $A \in \text{Equ}(\Delta, S)$ and let B be a matrix diagonally similar to A and such that B has nonnegative diagonal blocks in a (and therefore every) Frobenius

normal form. Since the diagonal blocks of B are clearly in Equ (Δ, N) it follows that $B \in$ Equ (Δ, N). Hence $A \in$ Equ (Δ, N) and (i) follows from (vi). □

THEOREM 6.14. *Let* $S \subseteq N$.

I. *If* $n \leqq 2$, *then the following are equivalent.*

(i) S *is sufficient,*

(ii) $\langle n \rangle \subseteq CD(S)$.

II. *If* $n \geqq 3$, *then* (i) *is equivalent to*

(iii) (a) $\{n-1, n\} \subseteq CD(S)$.

and

(b) $\langle n-1 \rangle \subseteq S$.

Proof. I: Let $n \leqq 2$.

(i) $\Rightarrow$ (ii). Since S is sufficient, it is also Δ-sufficient and the result follows from Theorem 6.12.

(ii) $\Rightarrow$ (i). Let $\langle n \rangle \subseteq CD(S)$. Suppose $A \in$ Equ (S). Then, by Lemma 6.8 all circuit products of A are positive. Since, for $i, j \in \langle n \rangle$ there is at most one nonempty simple path from i to j in $G(A)$, the conditions of Theorem 5.14, Part (v) are satisfied for all $i, j \in \langle n \rangle$. Hence, by Theorem 5.14, $A \in$ Equ (N) and the implication (ii) $\Rightarrow$ (i) follows.

II: (i) $\Rightarrow$ (iii), Part (a). With the same proof as in Part I, we have $\langle n \rangle \subseteq CD(S)$.

(i) $\Rightarrow$ (iii), Part (b). Let $2 \leqq k \leqq n-1$. To prove this implication it is enough to construct a matrix $B(k) \in$ Equ $(S)\backslash$Equ (N) if either $1 \notin S$ or $k \notin S$. We let the arc set of $G(B(k))$ consist of $1 \to k+1$, and $i \to i+1$, $i = 1, \cdots, k$. We define the $(1, k+1)$-element of $B(k)$ to be -1 and all other nonzero elements to be 1. For example, if $k = 2$ and $n = 4$, then

$$B(k) = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Let $i, j \in \langle n \rangle$. If either $1 \notin S$ or $k \notin S$ then there is at most one path from i to j in $G(B(k))$ whose length lies in S. Hence, by Lemma 4.6, we have $A \in$ Equ (S). But there are two paths from 1 to k in $G(B(k))$ whose corresponding products have different signs. Hence, again by Lemma 4.6, $A \notin$ Equ (N).

(iii) $\Rightarrow$ (i). Suppose that (iii) holds. Let $A \in$ Equ (S). Let $i, j \in \langle n \rangle$. Let α and β be simple paths in $G(A)$. Since $|\alpha| < n$, and $|\beta| < n$, we have by Lemma 4.6 that $\prod_\alpha(A) = \prod_\beta(A)$. Since $\langle n \rangle \subseteq CD(S)$, it follows from Lemma 6.8 that all circuit products of A are positive. Hence the conditions of Theorem 5.2, Part (v) are satisfied. By Theorem 5.2 we now obtain $A \in$ Equ (N) and (iii) $\Rightarrow$ (i) is proved. □

We note that, for $n \geqq 3$, neither of the conditions (iii)(a) or (iii)(b) of Theorem 6.14 alone implies that S is sufficient, or even that S is Δ-sufficient. This is clear from Theorem 6.12 since neither condition implies that $\langle n \rangle \subseteq CD(S)$.

COROLLARY 6.15. *Let* $n \geqq 3$. *Let* $S \subseteq N$.

I. *If* S *is sufficient then* $|S| \geqq n$.

II. *The following conditions are equivalent*:

(i) S *is sufficient and* $|S| = n$.

(ii) $S = \{1, \cdots, n-1, m\}$ *where* $n+1 \leqq m \leqq 2n-2$.

(iii) S *is optimal sufficient.*

Proof.

I. This is obvious by Theorem 6.14.

II. (i) $\Rightarrow$ (ii). By Theorem 6.14 we have $S = \{1, \cdots, n-1, m\}$. If $m = 0$ or

$m = n$ then $n \notin CD(S)$ and S is not sufficient by Theorem 6.14. Hence $m > n$. Suppose that $m > 2n - 2$. Then it follows that

$$D(S) = \{1, \cdots, n-2, m-n+1, \cdots, m-1\}. \tag{6.16}$$

Let $p, q \in D(S)$ where $p < q$. Then, by (6.16), either $p < n - 1$ or $q - p < n - 1$. Hence, gcd $(p, q) < n - 1$. It follows that gcd $(T) < n - 1$ for any subset T of $D(S)$ with $|T| > 1$. Since $n - 1 \notin D(S)$ and just one positive multiple of $n - 1$ belongs to $D(S)$ we also have $n - 1 \notin CD(S)$, which contradicts Theorem 6.14. The implication is now proved.

(ii) $\Rightarrow$ (iii). By Theorem 6.14, S is sufficient. The optimal sufficiency of S follows from Part I.

(iii) $\Rightarrow$ (i). Let S be an optimal sufficient set. Then clearly S is sufficient. By Theorem 6.14 the set $T = \{1, \cdots, n - 1, n + 1)$ is sufficient with $|T| = n$. Hence, by Part I we have $|S| = n$. □

We now use Corollary 6.15 to show that a minimal sufficient set is not necessarily an optimal sufficient set.

Example 6.17. Let $n \geqq 3$ and let $S = \{1, \cdots, n - 1, 2n - 1, 3n - 2\}$. Then $\langle n \rangle \subseteq CD(S)$ and so, by Theorem 6.14, S is sufficient. Let S' be a subset of S of cardinality n. Observe that S' cannot satisfy condition (ii) of Corollary 6.15. Hence, by Corollary 6.15, S' is not sufficient. Thus, S is a minimal sufficient set, but, again by Corollary 6.15, S is not an optimal sufficient set.

It is clear that our definitions and results raise a number of interesting questions. Some are purely number theoretic, others involve a mixture of matrix and number theory. A general problem is to characterize the $(\mathscr{A}, \Gamma)$-sufficient [minimal $(\mathscr{A}, \Gamma)$-sufficient, optimal $(\mathscr{A}, \Gamma)$-sufficient] sets for given $\mathscr{A} \subseteq \mathscr{U}$ and $\Gamma \subseteq \langle n \rangle \times \langle n \rangle$.

In view of Theorem 6.12 the following open questions are of interest.

Open Questions 6.18.

(i) Characterize subsets S of N such that $CD(S) \supseteq \langle n \rangle$.

(ii) Characterize subsets S of N which are minimal with respect to the property $CD(S) \supseteq \langle n \rangle$.

Remark 6.19. In Definition 4.4 the restriction to $A \in \mathscr{U}$ (viz. $A \in \mathbb{C}^{nn}$ such that $\rho(|A|) < 1$) and the use of power series with all coefficients equal to 1 are technicalities. Alternatively, we could have considered throughout arbitrary $A \in \mathbb{C}^{nn}$ and nonnegative sequences

$$\langle C \rangle = (c_1, c_2, \cdots\cdots\cdots)$$

such that $\sum_{s \in N} c_s |A|^s$ converges. In this approach one then defines the equality class Equ $(\mathscr{A}, \Gamma\ S)$ to consist of all $A \in \mathscr{A}$ such that for some nonnegative sequences $\langle C \rangle$ with $c_s \neq 0$ if and only if $s \in S$, $\sum_{s \in N} c_s |A|^s$ converges and

$$\left| \sum_{s \in N} (c_s A^s)_\Gamma \right| = \sum_{s \in N} (c_s |A|^s)_\Gamma.$$

Since the proof of our fundamental lemma, Lemma 4.6, is unchanged, our results go through to this more general situation and reduce to the previous results for $A \in \mathscr{U}$. The concept of sufficiency remains unchanged. We illustrate by means of an example.

Example 6.20. Let $n \leqq 10$. If $S = \{3, 9, 10, 13, 18\}$ then $CD(S) = \langle 10 \rangle \cup \{15\}$ and hence, by Theorem 6.12, S is $(\mathscr{J}, \langle n \rangle)$-sufficient. In other words, let A be an irreducible

$n \times n$ matrix, $n \leqq 10$, and let c_s be positive, $s = 3, 9, 10, 13, 18$. Then the equality

$$|c_3A^3 + c_9A^9 + c_{10}A^{10} + c_{13}A^{13} + c_{18}A^{18}|$$
$$= c_3|A|^3 + c_9|A|^9 + c_{10}|A|^{10} + c_{13}|A|^{13} + c_{18}|A|^{18}$$

implies that for all nonnegative d_s, $s \in N$, we have

$$\left| \sum_{s \in N} d_sA^s \right| = \sum_{s \in N} d_s|A|^s,$$

provided that the second series converges. In particular, if $\rho(|A|) < 1$, then

$$|(I - A)^{-1}| = (I - |A|)^{-1}.$$

REFERENCES

[1] P. M. COHN, *Universal Algebra*, Harper and Row, New York, 1965.

[2] G. M. ENGEL AND H. SCHNEIDER, *Cyclic and diagonal products on a matrix*, Linear Algebra Appl., 7 (1973), pp. 301–335.

[3] M. FIEDLER AND V. PTAK, *Cyclic products and an inequality for determinants*, Czechoslovak Math. J., 19 (1969), pp. 428–450.

[4] S. FRIEDLAND, D. HERSHKOWITZ, AND H. SCHNEIDER, *Matrices whose powers are M-matrices or Z-matrices*, Trans. Amer. Math. Soc., 300 (1987), pp. 343–366.

[5] D. HERSHKOWITZ AND H. SCHNEIDER, *On 2k-twisted graphs*, European J. Combin. (to appear).

[6] A. NEUMAIER, *The extremal case of some matrix inequalities*, Arch. Math., 43 (1984), pp. 137–141.

[7] A. OSTROWSKI, *Über die Determinanten mit überwiegender Hauptdiagonale*, Comment. Math. Helv., 10 (1937), pp. 69–96.

[8] B. D. SAUNDERS AND H. SCHNEIDER, *Flows on graphs applied to diagonal similarity and diagonal equivalence for matrices*, Discrete Math., 24 (1981), pp. 205–220.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 1, January 1988

SOME SIGN PATTERNS THAT PRECLUDE MATRIX STABILITY*

CLARK JEFFRIES† AND CHARLES R. JOHNSON‡

Abstract. The principal concern of this paper is with real matrices whose undirected graphs are trees. To better understand potential stability of sign pattern classes, two simple criteria are given that preclude stability throughout a sign pattern class. In addition, those sign patterns that preclude eigenvalues with real part equal to 0 are characterized and the constant inertia within such classes is determined. Such tests may be computationally significant, as calculations with specific matrices may be subject to round off error uncertainties.

Key words. potential stability, sign pattern matrix, stable matrix, tree graph

AMS(MOS) subject classifications. 47A20, 15A57, 93D99

1. Introduction.

The *inertia* of an n-by-n real matrix A is the triple

$$i(A) = (i_+(A), i_-(A), i_0(A))$$

in which $i_+(A)$ is the number of eigenvalues of A with positive real part, $i_-(A)$ the number with negative real part and $i_0(A)$ the number with zero real part—each counting any multiplicity; necessarily, $i_+(A) + i_-(A) + i_0(A) = n$. The matrix A is called *stable* if $i_-(A) = n$ because the equilibrium $\hat{x} = 0$ will be globally stable in the dynamical system $\dot{x} = Ax$ if and only if A is stable.

We are interested here in what may be concluded about the stability or instability of A purely from the $+$, $-$, 0 sign pattern of the entries of $A = (a_{ij})$. For this reason, we call an n-by-n matrix $B = (b_{ij})$ whose entries are chosen from among the symbols $\{+, 0, -\}$ a *sign pattern matrix*, and we identify with each sign pattern matrix the natural class of all real matrices $A = (a_{ij})$ such that $a_{ij} > 0$ (resp. $=$, < 0) if and only if $b_{ij} = +$ (resp. $= 0$, $= -$). Matrix operations with sign pattern matrices are carried out in the obvious way when unambiguous. For example, we call a diagonal sign pattern matrix none of whose diagonal entries is 0 a *signature matrix*, and left multiplication of a sign pattern matrix by a signature matrix uniformly affects the signs within each row.

A sign pattern matrix is called *sign stable* (respectively, *potentially stable*) [H], [Q], [B] if every (respectively, some) real matrix in the associated class is stable. The sign stable matrices have been characterized in [JKvdD], and several authors have discussed potential stability without any definite results thus far. Our interest here is in further understanding potentially stable sign patterns; however, our results are of a negative sort. We call a sign pattern matrix that is *not* potentially stable *sign unstable*; thus a sign pattern matrix is sign unstable if no matrix in the associated class is stable. Our goal is a characterization of certain sign unstable sign patterns. Clearly such sign patterns cannot be potentially stable.

By the (undirected) graph G of an n-by-n sign pattern matrix $B = (b_{ij})$, we mean a graph on vertices $1, 2, \cdots, n$ with an undirected edge between i and j if and only if b_{ij} *or* $b_{ji} \neq 0$. We concentrate here upon sign pattern matrices whose graphs are trees. Such a matrix is irreducible if and only if $a_{ij} \neq 0$ whenever $a_{ji} \neq 0$, and the eigenvalue possibilities within the class depend only upon the signs of the a_{ii} and of the products $a_{ij}a_{ji}$. We

* Received by the editors August 30, 1986; accepted for publication (in revised form) March 31, 1987. This work was supported in part by National Science Foundation grant DMS-8713762 and Office of Naval Research contracts N00014-86-K0012 and 0693.

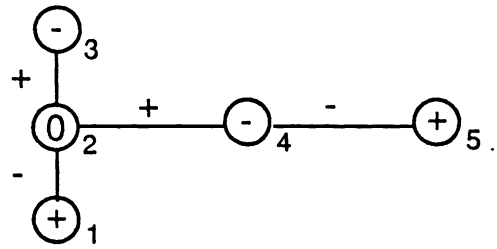
† La Ronge, Saskatchewan, Canada, and Mathematical Sciences Department, Clemson University, Clemson, South Carolina 29634-1907.

‡ Mathematics Department, College of William and Mary, Williamsburg, Virginia 23185.

describe in a natural way all relevant information about such a sign pattern class in a *signed* tree whose vertices may be signed +, −, or 0 and whose edges may be signed +, −. For example, we identify the sign pattern matrix

$$B = \begin{bmatrix} + & + & 0 & 0 & 0 \\ - & 0 & - & + & 0 \\ 0 & - & - & 0 & 0 \\ 0 & + & 0 & - & - \\ 0 & 0 & 0 & + & + \end{bmatrix}$$

with



Here, a + edge between i and j means that b_{ij} and b_{ji} are both + or both − and a − edge means that one is + and one is −; the sign of vertex i (+, −, or 0) is simply the sign of the i, i entry. If the sign is not zero, the vertex is called *distinguished.* We call an irreducible sign pattern matrix whose graph is a tree a *tree sign pattern* (t.s.p) matrix. As there is a one-to-one correspondence between t.s.p. matrices and signed trees, we shall use these interchangeably; we shall also move freely between concepts about graphs and matrices, when any ambiguity is benign. We call a t.s.p. matrix *symmetric* if each edge of the tree is + and *skew-symmetric* if each edge is −.

Two useful factorizations may be associated with each t.s.p. matrix. If B is an n-by-n t.s.p. matrix the *skew-symmetric factorization* of B is

$$B = S_1 B_1$$

in which S_1 is a signature matrix whose 1, 1 entry is + and B_1 is a skew-symmetric t.s.p. matrix. We call B *sign consistent* if not both + and − occur as diagonal entries in B_1. If B is sign consistent, let $\tilde{B}$ be obtained from B by sign consistently replacing all 0 diagonal entries with + or −. We call $\tilde{B}$ the sign completion of B. If any vertex in a sign consistent B is not signed 0, then $\tilde{B}$ is uniquely determined. Otherwise $\tilde{B}$ can be one of two t.s.p. matrices with opposite (+ and − interchanged) signs. In any $\tilde{B}$, nodes connected by a + [resp. −] edge are of opposite [same] sign.

The *symmetric factorization* of B is

$$B = S_2 B_2$$

in which S_2 is a signature matrix whose 1, 1 entry is +, and B_2 is symmetric t.s.p. matrix. Each factorization is unique, and the matrices S_i and B_i, $i = 1, 2$, are easily determined from B. It is a very open question to determine whether a t.s.p. matrix is potentially stable or sign unstable. All (irreducible) sign stable matrices have been classified and are t.s.p. matrices [JKvdD].

2. Sign instability tests. We present here two simple results which allow many t.s.p. matrices to be identified as sign unstable. For this we require two lemmas.

LEMMA 1. *If, for a given n-by-n matrix A, there exists a nonsingular n-by-n Hermitian matrix G such that*

$$GA = H + S$$

with H positive semidefinite Hermitian and S skew-Hermitian, then

$$i_+(A) \leqq i_+(G) \quad \text{and} \quad i_-(A) \leqq i_-(G).$$

Proof. If H is positive definite, then $i_0(A) = 0$, and this is the well-known equality of inertias result [CDJ]. In our case, choose $\varepsilon > 0$ sufficiently small so that the perturbed matrix $A + \varepsilon G^{-1}$ satisfies

$$i_+(A) \leqq i_+(A + \varepsilon G^{-1}) \quad \text{and} \quad i_-(A) \leqq i_-(A + \varepsilon G^{-1}).$$

However, we have

$$G(A + \varepsilon G^{-1}) = (H + \varepsilon I) + S$$

so that $i_+(A + \varepsilon G^{-1}) = i_+(G)$ and $i_-(A + \varepsilon G^{-1}) = i_-(G)$ because $H + \varepsilon I$ is positive definite. The asserted conclusions follow from these equalities and inequalities. □

LEMMA 2. *If A and B are n-by-n Hermitian and nonsingular, then* $i_+(BA) = 0$ *implies* $i_+(B) + i_+(A) = n$.

Proof. The proof is Corollary 2 of [J]. □

To apply these facts to our situation, we first note a familiar fact. If $A = (a_{ij})$ is an irreducible n-by-n matrix whose graph G is a tree, then there is a positive diagonal matrix D such that symmetrically placed off-diagonal entries of DA are the same in absolute value. It follows that if A is a real matrix in the class associated with a t.s.p. matrix B, then there is a factorization

$$A = D_1 A_1$$

in which D_1 is a nonsingular diagonal matrix and A_1 is a diagonal matrix plus a skew-Hermitian matrix and a factorization

$$A = D_2 A_2$$

in which D_2 is a nonsingular diagonal matrix and A_2 is symmetric. These correspond to the skew-symmetric and symmetric factorizations of B, and each is unique if the 1, 1 entry of D_i, $i = 1, 2$, is taken to be one.

Note that if S is a signature (sign pattern) matrix, $i(S)$ is well defined as the inertia of any matrix in the class associated with S.

The two results of this section are the following.

THEOREM 1. *Let* $B = S_1 B_1$ *be the skew-symmetric factorization of the n-by-n t.s.p. matrix B. If no diagonal entries of* B_1 *are* $-$, *then*

$$i_+(A) \leqq i_+(S_1) \quad \text{and} \quad i_-(A) \leqq i_-(S_1)$$

for all matrices A in the class associated with B. If no diagonal entries of B_1 *are* $+$, *then*

$$i_+(A) \leqq i_-(S_1) \quad \text{and} \quad i_-(A) \leqq i_+(S_1)$$

for all matrices A in the class associated with B.

Proof. The two conclusions are equivalent via replacement of B by $-B$. The first conclusion is an application of Lemma 1, as each A in the class associated with B may

be factored:

$$D_1^{-1}A = A_1$$

with $i(D_1^{-1}) = i(S)$ and $A_1 = E + T$, in which E is a positive semidefinite diagonal matrix and T is skew-symmetric. □

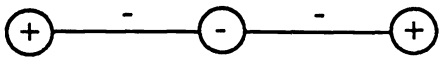
THEOREM 2. *Let $B = S_2B_2$ be the symmetric factorization of the n-by-n t.s.p. matrix B. If B is potentially stable, there is a symmetric matrix A_2 in the class associated with B_2 such that*

$$i_+(A_2) = n - i_+(S_2).$$

Proof. Lemma 2 may be applied using the same ideas as in the proof of Theorem 1. □

We illustrate the use of Theorems 1 and 2 to verify sign instability with some examples.

Example 1. Any t.s.p. matrix associated with the signed tree



is sign unstable using Theorem 2. For example, the symmetric factorization of

$$B = \begin{bmatrix} + & + & 0 \\ - & - & + \\ 0 & - & + \end{bmatrix}$$

is

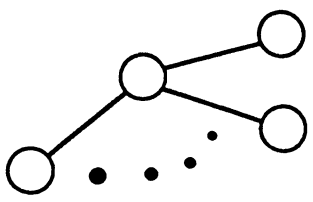
$$S_2B_2 = \begin{bmatrix} + & 0 & 0 \\ 0 & - & 0 \\ 0 & 0 & + \end{bmatrix}\begin{bmatrix} + & + & 0 \\ + & + & - \\ 0 & - & + \end{bmatrix}.$$

We notice that

$$\begin{bmatrix} + & 0 \\ 0 & + \end{bmatrix}$$

is the principal submatrix of B_2 in rows and columns 1 and 3, so $i_+(A_2) \geqq 2$. Also $i_+(S_2) = 2$. Hence, the necessary condition of Theorem 2 cannot be met, and B cannot be potentially stable. In this particular case it is quite complicated to verify that B_2 is sign unstable by direct calculation. This example illustrates a general way in which Theorem 2 may be applied.

If a lower bound on $i_+(A_2)$ may be found which is greater than $n - i_+(S_2)$ (e.g., if we extract a large principal submatrix of B_2 with obvious inertia and realize that this bounds the inertia of A_2 because of the interlacing inequalities), then Theorem 2 implies that B is sign unstable. Since a tree is bipartite, a diagonal principal submatrix of size at least $\frac{1}{2}n$ is always available. In case of the graph



a diagonal principal submatrix of size $(n - 1)$ is available.

Example 2. If

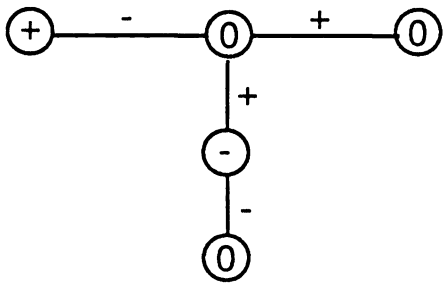
$$B = \begin{bmatrix} - & + & + & + & + & + & + & + \\ + & + & & & & & & \\ - & & + & & & & & \\ + & & & + & & & & \\ - & & & & + & & & \\ + & & & & & - & & \\ - & & & & & & - & \\ + & & & & & & & + \end{bmatrix},$$

then

$$B = \begin{bmatrix} + & & & & & & & \\ & + & & & & & & \\ & & - & & & & & \\ & & & + & & & & \\ & & & & - & & & \\ & & & & & + & & \\ & & & & & & - & \\ & & & & & & & + \end{bmatrix} \begin{bmatrix} - & + & + & + & + & + & + & + \\ + & + & & & & & & \\ + & & - & & & & & \\ + & & & + & & & & \\ + & & & & - & & & \\ + & & & & & - & & \\ + & & & & & & + & \\ + & & & & & & & + \end{bmatrix}$$

is the symmetric factorization of B. Since $i_+(S_2) = 5$ and $i_+(A_2) \geqq 4$, B is sign unstable.

Example 3. The t.s.p. matrices associated with the signed tree



are sign unstable as may be seen from Theorem 1 but not Theorem 2. The skew-symmetric factorization of

$$B = \begin{bmatrix} + & + & & & \\ - & 0 & + & + & \\ & + & 0 & & \\ & + & & - & + \\ & & & - & 0 \end{bmatrix}$$

is

$$\begin{bmatrix} + & & & \\ & + & & \\ & & - & \\ & & & - \end{bmatrix} \begin{bmatrix} + & + & & & \\ - & 0 & + & + & \\ & - & 0 & & \\ & - & & + & - \\ & & & + & 0 \end{bmatrix} = S_1 B_1.$$

Thus B is sign consistent and by Theorem 1, $i_-(A) \leqq 3$ for any A in the class associated with B (actually $i(A) = (2, 3, 0)$ for such an A), and B is sign unstable. The symmetric

factorization of B, however, is

$$\begin{bmatrix} + & & & & \\ & - & & & \\ & & - & & \\ & & & - & \\ & & & & + \end{bmatrix} \begin{bmatrix} + & + & & & \\ + & 0 & - & - & \\ & - & 0 & & \\ & - & & + & - \\ & & & - & 0 \end{bmatrix} = S_2 B_2.$$

Since the determinant of any matrix A_2 in the class associated with B_2 is positive, $i(A_2) = (3, 2, 0)$ for any such matrix. As $i_+(S_2) + i_+(A_2) = 5$, Theorem 2 does not preclude the (potential) stability of B as Theorem 1 did. It should be noted that the potential stability of the matrix in Example 1 is not precluded by Theorem 1 as it is by Theorem 2.

3. Sign pattern matrices with constant inertia. We recall from [JvdD] two color tests. In the 0-color test (read "zero color test") we color each vertex of the (tree) graph of an irreducible matrix black or white so that

(i) no black vertex is a neighbor of exactly one white vertex;

(ii) each maximal white block as a subgraph is either: a single undistinguished vertex; or a subgraph which has at least 2 vertices, which has each end vertex distinguished, and which is not sign consistent.

We define an Im-coloring of vertices of G to be again a scheme for coloring each vertex black or white so that condition (i) is fulfilled as well as

(ii) each maximal white block as a subgraph of G contains at least one "$-$" edge and is not sign consistent.

The dynamical system $\dot{x} = Ax$ admits a constant (resp. sinusoidal) trajectory if and only if some A in the sign pattern class has 0 (resp. $\sqrt{-1}$) as eigenvalue if and only if G admits a 0-coloring (resp. Im-coloring) with at least one white vertex [JvdD, Thms. 3 and 5].

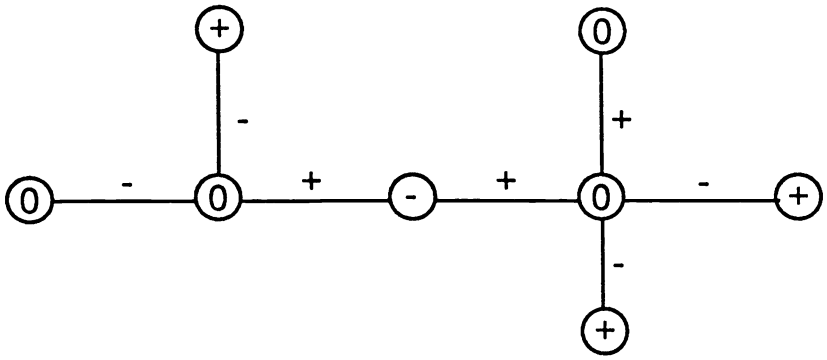
THEOREM 3. *Suppose $n > 1$ and G is a tree graph of an irreducible sign pattern matrix. Then there is only one* 0*-coloring (all black) and only one* Im*-coloring (all black) for G if and only if $i_0(A) = 0$ for any matrix A of the given sign pattern. (In such case $i_+(A)$ and $i_-(A)$ are necessarily constants for all such A.)*

Proof. Let us identify in a natural way $n \times n$ matrices and n^2-dimensional space. Consider a continuous curve in n^2-dimensional space lying in the cone of all matrices of the given sign pattern. To each point on the curve are associated the n eigenvalues of the corresponding $n \times n$ matrix. The fundamental theorem of algebra, the fact that the determinant function is continuous, and the fact that the zeros of a polynomial depend continuously on its coefficients together imply that the eigenvalues of a matrix on the curve move about continuously in the complex plane. In particular, since the theorems in [JvdD] preclude the occurrence of any eigenvalues on the imaginary axis, the number of eigenvalues with positive real parts is conserved throughout the cone. Of course, the same can be said of eigenvalues with negative real parts. A corollary of the same theorems is that if inertia is constant with $i_0(A) = 0$ throughout the cone, then the only colorings are all black.

Consider a matrix $A = (a_{ij})$ in the cone having entries of large magnitude and a matrix B, the graph of which is the sign completion of G. Suppose, in fact, that $b_{ij} = a_{ij}$ if $a_{ij} \neq 0$ and $0 < |b_{ii}| < \varepsilon$ if $a_{ii} = 0$. If ε is suitably small then the eigenvalues of B are arbitrarily close to the eigenvalues of A. In particular, the inertia of B can be assumed to be the inertia of A.

Since inertia is conserved throughout the cone in n^2-dimensional space of matrices of the sign pattern of B, choose such a matrix $\tilde{B}$ having $(\tilde{b}_{ij}) < \delta$ for $i \neq j$ and $|b_{ii}| = 1$. For δ suitably small, the characteristic polynomial of $\tilde{B}$ is approximately $(x-1)^{i_+(A)}(x+1)^{i_-(A)}$. □

Theorem 3 can be applied to the graph



to show any matrix of the sign pattern has inertia (6, 2, 0). That is, the graph is sign consistent and so every subgraph is sign consistent. The color tests are therefore not difficult to check. The sign completion of the graph has six "+" vertices and two "−" vertices.

Authors' note. Based, in part, upon the results of this paper, T. Summers has been classifying t.s.p. matrices with regard to potential stability in hopes of gaining insight into the general problem. A summary of the results is available from C. Johnson.

REFERENCES

[B] T. BONE, *Qualitative stability properties of matrices*, Ph.D. thesis, Univ. of Washington, Seattle, Washington, 1985.

[CDJ] D. CARLSON, B. DATTA AND C. R. JOHNSON, *A semi-definite Lyapunov theorem and the characterization of tridiagonal D-stable matrices*, SIAM J. Algebraic Discrete Methods, 3 (1982), pp. 293–304.

[H] W. HAHN, *Stability of Motion*, Springer–Verlag, New York, 1967.

[JKvdD] C. JEFFRIES, V. KLEE AND P. VAN DEN DRIESSCHE, *When is a matrix sign stable*? Canad. J. Math., 29 (1977), pp. 315–326.

[JvdD] C. JEFFRIES AND P. VAN DEN DRIESSCHE, *Eigenvalues of matrices with tree graphs*, to appear.

[J] C. R. JOHNSON, *The inertia of a product of two hermitian matrices*, J. Math. Anal. Appl., 57 (1977), pp. 85–90.

[Q] J. QUIRK, *The correspondence principle: a macroeconomic application*, Internat. Econom. Rev., 9 (1968), pp. 294–306.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 1, January 1988

A TREE MODEL FOR SPARSE SYMMETRIC INDEFINITE MATRIX FACTORIZATION*

JOSEPH W. H. LIU†

Abstract. A tree model is presented to study the sparse factorization of large symmetric indefinite matrices by the diagonal pivoting method. The basic structure uses the elimination tree of symmetric matrices and the notion of delayed elimination. The factorization process for indefinite systems can be viewed as a sequence of tree transformations based on both the structural information and numerical data values. This provides a model as a common basis to study various numerical aspects of sparse symmetric indefinite decomposition.

Key words. sparse matrix, indefinite symmetric matrix, tree model, diagonal pivoting, elimination tree, delayed elimination

AMS(MOS) subject classifications. 65F50, 65F25

1. Introduction. In this paper, we study the *diagonal pivoting method* [4] in factoring large sparse symmetric indefinite matrices. The method uses a mixture of 1×1 and 2×2 pivots to produce an LBL^T decomposition, where the factor matrix L is unit lower triangular and B is block diagonal with blocks of size of either 1 or 2.

The method will be considered in connection with the so-called elimination tree structure [13], [17], which is defined for each sparse symmetric matrix structure. The elimination tree represents a class of ideal elimination sequences if we assume that no pivoting for numerical stability is necessary. Our approach is to use the tree structure as a pivot selection guide, so that even with the added stability requirement, the sequence of stable pivots selected will form an elimination tree that deviates as little as possible from the original one.

In this paper, we consider the notion of delayed elimination in sparse decomposition. In the dense case [2], [11], when row/column j is considered not suitable for elimination as a 1×1 pivot, some later row/column will be moved forward to form a 2×2 pivot with j. However, in the sparse case, it is more appropriate to delay the elimination of row/column j to a later stage. The notion of delayed elimination first appears in the multifrontal work by Duff and Reid [7]. Our treatment here helps to bring out the important role of this idea in the context of sparse symmetric factorization. Furthermore, we provide some quantitative bounds on the impact of delayed elimination on fills in the resulting triangular factors.

A tree model can be formulated according to the use of delayed elimination on an elimination tree. This model provides a systematic view of the elimination process. At each step, pivots can only be selected from the nodes in a specific subtree, which represents the set of preferred candidates. Nodes from this subtree will incur the least amount of structural damage if selected as pivots. The actual pivots selected depend on the partial pivoting strategy and the numerical values of the matrix.

This tree model plays an important role in symmetric factorization. It provides a better understanding in the choice of 2×2 pivots and it helps to reveal the fundamental importance of delayed elimination. On the basis of the model, researchers can focus

* Received by the editors October 3, 1986; accepted for publication (in revised form) April 21, 1987. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant A5509, 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., and by the U.S. Air Force Office of Scientific Research under contract AFOSR-ISSA-85-0083.

† Department of Computer Science, York University, North York, Ontario, Canada M3J 1P3.

more on other algorithmic aspects of sparse symmetric factorization, especially on the design of data structures and the numerical computations.

The reader is assumed to be familiar with the graph-theoretic terminology associated with sparse matrix computation: adjacent set, subgraph, fill, ordering, elimination graph and other related concepts. All the necessary material can be found in [10]. Moreover, notions related to tree structures are also assumed: parent/child nodes, ancestor/descendant nodes, root, paths, subtrees. The reader is referred to [1].

An outline of this paper follows. In § 2, we provide a brief overview of background material in symmetric matrix factorization. In particular, we review the diagonal pivoting method for indefinite matrices and consider the elimination graph model in this context. The elimination tree structure is also defined and some relevant properties are stated.

Section 3 considers the impact on the structure of a given elimination tree due to relabeling. We establish the observation that any relabeling within a subtree will not affect parts that are outside this subtree. This leads to the notion of delayed elimination. An upper bound on possible fill increase due to delayed elimination is given.

The tree model is described in § 4. The entire elimination process can be viewed as a sequence of tree transformations starting with the elimination tree. At each step, the tree provides the structural information necessary to guide the selection of the next pivot. Moreover, each transformation is a simple tree manipulation function.

In § 5, we relate the tree model to the multifrontal scheme of Duff and Reid [7]. The multifrontal method can be considered as one way of implementing the tree model. There are other ways depending on the data storage scheme, pivoting strategy, and numerical computation method. To substantiate this observation, we provide a different and new sparse factorization scheme for indefinite systems based on the tree model. Section 6 contains our concluding remarks.

2. Background on symmetric matrix factorization.

2.1. Diagonal pivoting method for indefinite matrices. In this paper, we employ the diagonal pivoting method [2]–[4] in the symmetric factorization of sparse indefinite matrices. The method is a variant of symmetric Gaussian elimination, wherein pivots are always taken from the diagonal but they may be of order 1 or 2. With an appropriately chosen pivoting strategy, the method is known to be nearly as stable as conventional Gaussian elimination with pivoting. However, symmetry can now be exploited through the use of 2×2 block pivots.

There are many appropriate ways to select stable pivots for elimination. The one by Bunch and Parlett [4] can be viewed as a complete/total pivoting strategy. The later ones by Bunch and Kaufman [2], Dax [5], and Fletcher [9] can all be classified as methods using partial pivoting.

In [2], Bunch and Kaufman provide a number of partial pivoting strategies tailored for this approach. In particular, Algorithm D [2, pp. 169–170] seems to be most appropriate for sparse matrices. Indeed, the authors point out that "whenever a 1×1 pivot is used in Algorithm D, no interchanges are performed, which means . . . fewer opportunities to interfere with the structure of the system." The essence of the diagonal block pivoting approach using Algorithm D can be expressed algorithmically as follows:

```
for j := 1 to n do
  if column j has not been eliminated then
    begin
      if column j is a suitable 1 × 1 pivot then
          eliminate column j
```

```
    else
      begin
        find a column i (i > j) such that columns i and j form
            a suitable 2 × 2 pivot;
        eliminate columns i and j together
      end
  end;
```

In [8], Duff et al. consider the use of block diagonal pivots in the factorization of large sparse indefinite matrices. They recommend a partial pivoting strategy geared for sparse systems. In [7], Duff and Reid combine this pivoting strategy with the multifrontal approach to devise a very effective scheme for sparse indefinite matrix factorization. The author [12] provides a simple improvement to their pivoting strategy.

2.2. Elimination graph model for diagonal pivoting. From the pioneering work of Parter [15] and Rose [16], the symmetric factorization of large sparse positive-definite matrices can be conveniently studied by the *elimination graph model.* The factorization process can be viewed as generating a sequence of elimination graphs, each reflecting the structure of an intermediate matrix to be factored. For more details, the reader is referred to [10].

The basis of the model is the rule for transforming the elimination graphs. Let $G = G(A)$ be the undirected graph associated with a given sparse symmetric matrix A, and x be a node in G. Consider the elimination of the node x. We obtain the resulting elimination graph from G by deleting the node x and its incident edges, and making the nodes adjacent to x into a clique (or complete subgraph). Since we are dealing with a possible mixture of 1×1 and 2×2 pivots in symmetric indefinite factorization, let us first extend the elimination graph transformation to allow for *block elimination.*

Let K be a connected subgraph of G. Consider the elimination of nodes in K from G. It is easy to see that the resulting elimination graph can be obtained from G by deleting the subgraph K and edges connecting nodes in K to $G - K$, and making the nodes adjacent to K into a clique. Note that the transformed elimination graph is independent of the order in which the nodes in K are eliminated. Our context of using 2×2 pivots corresponds to the case where the connected subgraph K has exactly two nodes. It should be mentioned that the resulting elimination subgraph has the same structure whether two consecutive columns/vertices are eliminated individually as two 1×1 pivots or together as a 2×2 block pivot.

Therefore, for a given elimination sequence of node subsets (of size either 1 or 2) in block diagonal pivoting, the associated elimination graph sequence can be generated easily. However, this graph sequence is of little use in practice, since the node subset sequence is not known a priori. The choice of 1×1 or 2×2 block pivots depends on the numerical values of the matrix under consideration, and they cannot be determined with only the sparsity structure of the matrix.

2.3. Elimination tree structure for sparse matrices. One of the key structures in the study of symmetric sparse Cholesky factorization is the *elimination tree* [13], [17]. For a sparse symmetric matrix with a given row/column ordering, the elimination tree structure can be used to determine a class of orderings that are equivalent in terms of fills and operations. We shall be using the structure of an elimination tree to study block diagonal pivoting for indefinite matrices.

In this subsection, we provide a brief review on this tree structure. Let A be a given $n \times n$ sparse symmetric matrix. Consider the numerical symmetric LDL^T decomposition of A (see, for example, [11, Chap. 5]). It is well known that if A is indefinite, this numerical decomposition can be unstable. Furthermore, for certain nonsingular A, such a factori-

zation may not even exist. We shall use the notation $L[A]$ to represent the numerical triangular factor (if it exists) of A in the symmetric decomposition. When the matrix A is clear from context, L will be used.

In spite of the numerical shortcoming in factoring general symmetric A into LDL^T, it is still meaningful to consider the *structural* symmetric factorization of the structure of A. In the literature, there are existing efficient algorithms and robust implementations which will determine the structure of the triangular factor using the structure of A. Often, this is referred to as the *symbolic factorization* process [10].

Assume that all diagonal entries of A are logically nonzero and no pivoting is performed in the structural decomposition. Let $\hat{L}[A]$ be the structural triangular factor of A. Again, if A is clear from context, $\hat{L}$ will be used. We can now introduce the elimination tree in terms of $\hat{L}[A]$.

We define the elimination tree $T(A)$ of A to be the tree with n nodes $\{x_1, x_2, \cdots, x_n\}$, where node x_i is the parent of node x_j if and only if

$$i = \min \{r > j | \hat{L}_{rj} \neq 0\},$$

that is, if i is the row subscript of the first off-diagonal nonzero in column j of $\hat{L}$. Here, each node x_i is associated with row/column i of the matrix. We further assume that the matrix A is *irreducible*, so that the structure is indeed a tree, and x_n is the root of this tree. (If A is reducible, then the elimination tree defined above is actually a forest consisting of several trees.)

Figure 2.1 contains an 8×8 symmetric matrix structure A. The diagonal entries are labeled with their corresponding equation/variable numbers. Note that this matrix suffers two fills at locations (2, 8) and (6, 7), and each fill is depicted by an "O" in the matrix structure $\hat{L}[A]$ in the figure. The corresponding undirected graph and elimination tree is displayed in Fig. 2.2. This matrix and tree structure will be referred to throughout the remainder of this paper.

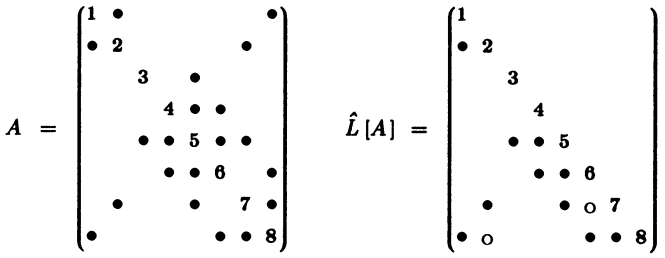


FIG. 2.1. *A matrix example and its structural triangular factor.*

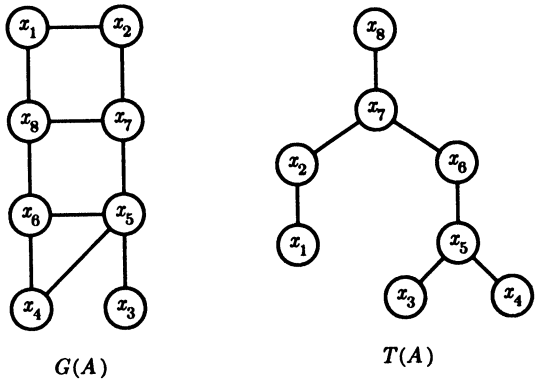


FIG. 2.2. *The graph $G(A)$ and elimination tree $T(A)$ of matrix A in Fig. 2.1.*

We introduce a *depth* function [1] here to be used later in the next section. For the root x_n, we define

$$\text{depth } (x_n) = 0.$$

For any node x_j ($j < n$), its depth value is defined to be

$$\text{depth } (x_j) = \text{depth } (x_p) + 1,$$

where x_p is the parent node of x_j in the elimination tree. For example, in Fig. 2.2, we have depth $(x_6) = 2$, and both depth (x_3) and depth (x_4) are 4. Clearly, depth (x_j) is the length of the path from x_j to the root x_n.

We also introduce the subtree notation to facilitate future discussion. Let y be a node in the elimination tree T. We shall use $T[y]$ to refer to the subtree rooted at the node y in the elimination tree. Moreover, $T[y]$ will also be used to refer to the set of nodes in this subtree. For example, in Fig. 2.2, the subtree $T[x_6]$ contains the node subset $\{x_3, x_4, x_5, x_6\}$.

We shall now state some properties of the elimination tree that are relevant to our study of indefinite factorization.

Observation 2.1. Any reordering that numbers child nodes before parent nodes in the elimination tree is equivalent to the original ordering.

In other words, the number of fills and the number of arithmetic operations to perform the factorization remain unchanged. Such orderings are referred to as *topological orderings* of the tree in the literature [19]. The tree structure provides some degree of flexibility in terms of the node elimination sequence without affecting the amount of fills and computation.

Observation 2.2. [13] For $i > j$, $\hat{L}_{ij}$ is nonzero if and only if $x_i \in \text{Adj } (T[x_j])$, where the Adj operator is taken in the graph $G(A)$.

By Observation 2.2, the number of nonzeros in the jth column of L is given by $|\text{Adj } (T[x_j])| + 1$. It should be noted that each $T[x_j]$, as a subgraph, is connected in the original graph $G(A)$. Furthermore, each node x_i adjacent to $T[x_j]$ is an ancestor node of x_j in the elimination tree.

Our strategy in using the elimination tree for sparse symmetric indefinite matrix factorization is as follows. From the symmetric matrix structure, determine its elimination tree. This tree structure will define a class of (equivalent) ideal orderings for fill and operation reduction, taking only the structure of the matrix into account. We shall use the tree structure as a *pivot selection guide*, so that any necessary reordering due to the use of 2×2 pivots for numerical stability reasons will deviate as little as possible from this tree.

3. Structural changes to the elimination tree.

3.1. Subtree relabeling. In practice, the block elimination sequence from the diagonal pivoting method is not known a priori. However, based on the structural information of the symmetric matrix, a fill-reducing node sequence (and hence its elimination tree) can be determined if we assume that no pivoting is necessary. The tree represents a class of ideal orderings for sparse elimination without taking the numerical values into consideration. In this section, we investigate the impact of relabeling on the structure of the elimination tree, when a rearrangement of node sequence is performed to obtain suitable 1×1 and 2×2 numerical pivots.

Observation 3.1. Consider a given subtree $T[y]$. If the nodes in $T[y]$ are to be ordered before the ancestors of y (that is, nodes on the path from y to the root), any relabeling of nodes in $T[y]$ will not alter the structure of the elimination tree associated

with nodes not in $T[y]$ (assuming that the relative order of nodes outside $T[y]$ remains unchanged).

This observation is key to the study of relabeling strategy due to stability requirement. It says that any renumbering of nodes in a subtree will not incur any structural damage to the remaining part of the elimination tree. In other words, the structural change is only local to the subtree involved. For example, consider the matrix A in Fig. 2.1. No matter how we rearrange symmetrically rows/columns 3, 4 and 5 of A among themselves, the corresponding elimination tree of any such renumbering will only affect the subtree $T[x_5]$. There is no structural change in the part of the tree outside this subtree, namely the part involving the nodes $\{x_1, x_2, x_6, x_7, x_8\}$. This structural preservation is important in terms of fills because of Observation 2.2.

3.2. Delayed elimination. In order to minimize structural changes to a given elimination tree, the previous subsection offers the observation that any relabeling should be confined locally to subtrees. What we discuss now is the actual relabeling strategy within a subtree.

Consider Algorithm D of Bunch and Kaufman [2]; an algorithmic version is given in § 2.1. If row/column j is viewed as an inappropriate 1×1 pivot, a later row/column i is determined and brought forward to be eliminated with j as a 2×2 pivot. In other words, row/column i is eliminated earlier than as scheduled. We shall refer to this as *advanced elimination.* In all other block pivoting algorithms for dense systems, the relabeling strategies use some form of advanced elimination. This is a satisfactory scheme for dense matrices since the reordered matrix structure remains unchanged.

However, it may be undesirable for sparse systems due to possible structural damage from the relabeling. Figure 3.1 gives a 6×6 matrix example, where "ε" is used to denote a numerical value much smaller than normal values indicated by "●". Advanced elimination will bring row/column 6 forward to go with row/column 1 as a 2×2 pivot. This obviously will cause severe fill-in.

The other alternative is the use of *delayed elimination.* This notion is implicit in the multifrontal scheme for indefinite sparse systems by Duff and Reid [7]. This means when a node (row/column) is deemed as inappropriate for a 1×1 pivot, its elimination will be delayed. For the example in Fig. 3.1, if we delay the elimination of the first row/column until after the last node, we obtain a much more desirable elimination sequence. In this case, no fill will occur.

We now consider the effect of delayed elimination on the structure of the elimination tree. Let A be the given symmetric matrix having

$$x_1, \cdots, x_k, \cdots, x_j, \cdots, x_n$$

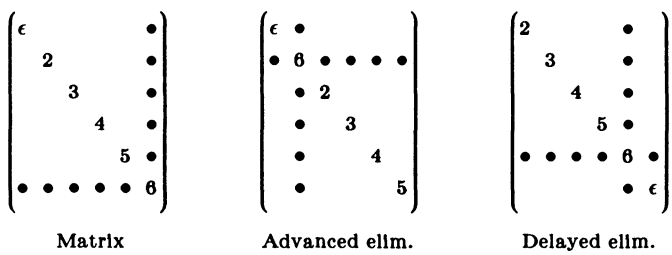


FIG. 3.1. *Advanced and delayed elimination on a matrix example.*

as its node elimination sequence, with $k < j$. Consider delaying the elimination of node x_k until immediately after x_j, so that the new node elimination sequence is given by

$$x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_j, x_k, x_{j+1}, \cdots, x_n.$$

Let the correspondingly permuted matrix be $\bar{A} = PAP^T$, which is obtained by moving the kth row/column to be after the jth row/column of A. That is, the jth row and column of $\bar{A}$ is the same as the kth row and column of A.

Observation 3.2. To delay the elimination of the node x_k immediately after x_j is the same as delaying the elimination of x_k to be immediately after the last ancestor of x_k before and including x_j.

By this observation, when we study the delayed elimination of a node x_k after x_j, it is sufficient to consider the case where x_j is an ancestor node of x_k, that is, $x_k \in T[x_j]$. We shall assume this in the remainder of this section.

Let T be the elimination tree of A, and $\bar{T}$ be that of $\bar{A} = PAP^T$. We shall provide some observations on the structural change from T to $\bar{T}$ as a result of the delayed elimination of the node x_k. The proofs are quite simple and they are omitted.

Observation 3.3. If x_c is a node *not* on the path from x_k to x_j, then $\bar{T}[x_c]$ and $T[x_c]$ are identical as tree structures (and hence are the same as node subsets).

Observation 3.4. As node subsets, we have $\bar{T}[x_k] = T[x_j]$.

Observation 3.5. If x_c is a node on the path from x_k to x_j other than x_k, then we have

(a) as node subsets, $\bar{T}[x_c] \subseteq T[x_c] - \{x_k\}$;
(b) $x_k \in \text{Adj}\,(\bar{T}[x_c])$;
(c) $\text{Adj}\,(\bar{T}[x_c]) \subseteq \text{Adj}\,(T[x_c]) \cup \{x_k\}$.

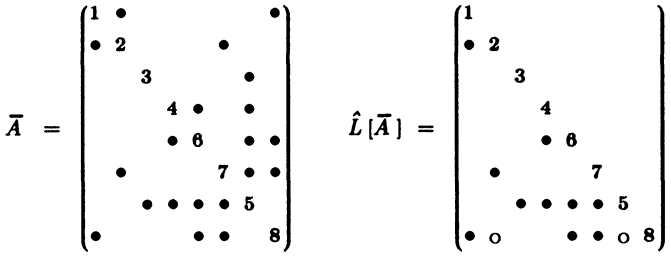


FIG. 3.2. *A permuted matrix structure of Fig.* 2.1.

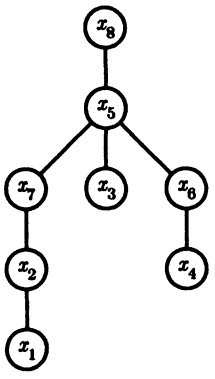


FIG. 3.3. *The elimination tree $\bar{T}$ of $\bar{A}$ in Fig.* 3.2.

To illustrate the structural change, we again use the matrix example in Fig. 2.1. Consider the delayed elimination of the node x_5 to be after x_7. The corresponding reordered matrix $\bar{A}$ and its triangular structure is given in Fig. 3.2. The elimination tree $\bar{T}$ for the permuted matrix is provided in Fig. 3.3. It is clear that the subtrees $T[x_5]$, $T[x_6]$ and $T[x_7]$ are the only subtrees in Fig. 2.2 with their node subsets changed in $\bar{T}$. The node subsets in both $\bar{T}[x_6]$ and $\bar{T}[x_7]$ are shrunk from $T[x_6]$ and $T[x_7]$, respectively. Moreover, $\bar{T}[x_5]$ is the only subtree in $\bar{T}$ that has its node subset enlarged from that of T.

3.3. Fill increase due to delayed elimination. We now consider the impact of delayed elimination on the number of fills in the triangular factor matrix. As before, we are delaying the elimination of x_k until immediately after its ancestor x_j. We shall use the notation $\eta(\blacksquare)$ to represent the number of nonzeros in "$\blacksquare$", which is either a vector or a matrix. For notational convenience, let $V = \hat{L}[A]$, and $W = \hat{L}[PAP^T]$.

LEMMA 3.6. *For* $c = 1, \cdots, k-1$ *and* $j+1, \cdots, n$,

$$\eta(W_{*c}) = \eta(V_{*c}).$$

Proof. It follows directly from Observations 3.3 and 2.2. □

LEMMA 3.7. $\eta(W_{*j}) = \eta(V_{*j})$.

Proof. By Observation 2.2, we have

$$\eta(W_{*j}) = |\text{Adj}\,(\bar{T}[x_k])| + 1$$

and

$$\eta(V_{*j}) = |\text{Adj}\,(T[x_j])| + 1,$$

and hence they are the same by Observation 3.4. □

LEMMA 3.8. *For* $c = k+1, \cdots, j$, *if* x_c *is on the path from* x_k *to* x_j, *then*

$$\eta(W_{*c-1}) \leqq \eta(V_{*c}) + 1;$$

otherwise,

$$\eta(W_{*c-1}) = \eta(V_{*c}).$$

Proof. If x_c is not on the path, then $\eta(W_{*c-1}) = |\text{Adj}\,(\bar{T}(x_c))| + 1$, since the node x_c is labeled $c-1$ in the new ordering. By Observation 3.3, this value is the same as $\eta(V_{*c})$. On the other hand, if x_c is on this path, the result follows from Observation 3.5. □

THEOREM 3.9. *With the new ordering, the number of fills in the structural matrix factor will be increased by no more than*

$$\text{depth}\,(x_k) - \text{depth}\,(x_j) + \eta(V_{*j}) - \eta(V_{*k}).$$

Proof. From the definition of depth, note that depth (x_k) − depth (x_j) is simply the number of nodes along the path from x_k to x_j not counting x_k. Combining Lemmas 3.6–3.8, we have

$$\begin{aligned}
\eta(W) &= \sum_{c=1}^{n} \eta(W_{*c}) \\
&= \sum_{c \neq j} \eta(W_{*c}) + \eta(W_{*j}) \\
&\leqq \sum_{c \neq k} \eta(V_{*c}) + \text{depth}\,(x_k) - \text{depth}\,(x_j) + \eta(V_{*j}) \\
&= \eta(V) + \text{depth}\,(x_k) - \text{depth}\,(x_j) + \eta(V_{*j}) - \eta(V_{*k}).
\end{aligned}$$

□

COROLLARY 3.10. *The increased number of fills due to the delayed elimination of the node x_k is always less than $n-k$.*

Proof. The result follows from Theorem 3.9 and the fact that

$$\text{depth}\,(x_k)-\text{depth}\,(x_j)\leqq j-k,$$

$$\eta(V_{*j})\leqq n-j+1,$$

$$\eta(V_{*k})>1. \qquad \square$$

The actual number of fills increased due to delayed elimination depends on the matrix structure, and in practice, it is usually quite modest. For example, the matrix in Fig. 3.3 is obtained by delaying the elimination of the node x_5 to after x_7 in Fig. 2.1. Then, by Theorem 3.9, the number of fills increased is bounded by

$$\text{depth}\,(x_5)-\text{depth}\,(x_7)+\eta(\hat{L}_{*7})-\eta(\hat{L}_{*5})=3-1+2-3=1,$$

and in this case, there is actually no increase in fills.

It should be noted that the quantity

$$\text{depth}\,(x_k)-\text{depth}\,(x_j)+\eta(V_{*j})-\eta(V_{*k})$$

is the *exact* difference in number of nonzeros between the jth row/column of W and the kth row/column of the original factor V. However, this quantity represents only an upper bound on the actual increased number of fills of the entire factor matrix as given in Theorem 3.9. It is due to possible reduction in fills in other rows/columns.

It is interesting to point out that moving a column forward to eliminate in advanced elimination can be treated as a sequence of delayed eliminations. Indeed, to eliminate x_j before x_k is the same as delaying the columns associated with nodes x_c, from $c=j-1, j-2, \cdots, k$ (in decreasing order) to be after x_j. Therefore, the increased number of fills for advanced elimination is potentially much greater than that of delayed elimination.

4. The tree model. Let A be the given sparse symmetric matrix. We assume further that the matrix A has been ordered to reduce fills. In this section, we consider the use of the elimination tree structure to generate a stable block elimination sequence with 1×1 and 2×2 pivots based on the numerical values in the given matrix. Let $x_1, x_2, \cdots, x_n$ be a given node elimination sequence on (the structure of) the matrix A, and $T_1 = T(A)$ be the corresponding elimination tree. The following algorithm uses a sequence of tree structures $T_1, T_2, \cdots, T_n$ to determine a block sequence.

ALGORITHM 4.1 (Block elimination sequence).

```
begin
  T_1 := T(A);
  for j := 1 to n do
    begin
      if there is a suitable 2 × 2 pivot using x_j and x_k for some x_k ∈ T_j[x_j]
        then eliminate {x_j, x_k} and transform the tree T_j to T_{j+1}
        else
          begin
            if x_j is a suitable 1 × 1 pivot
              then eliminate x_j and transform the tree T_j to T_{j+1}
              else set T_{j+1} := T_j   /* delay elimination of x_j */
          end
    end;
```

eliminate the remaining nodes in the tree T_{n+1}.
end.

In this algorithm, we use the notation $T_j[x_j]$ to denote the subtree rooted at the node x_j in the tree T_j. The tree T_j contains nodes that have not been eliminated. In particular, $T_j[x_j]$ represents the set of nodes in $T[x_j]$ whose elimination has been thus far delayed (except x_j). In the algorithm, preference is given to 2×2 pivots over 1×1 pivots if the 2×2 involves a node that was delayed earlier on.

As in Algorithm 2.1, we have left the numerical conditions for suitable 1×1 and 2×2 pivots unspecified for the time being. However, we need to provide the *tree transformation* rules for

$$T_1 \rightarrow T_2 \rightarrow \cdots \rightarrow T_n \rightarrow T_{n+1}.$$

This is important since the domain to search for suitable 2×2 pivots in step j is given by the subtree $T_j[x_j]$. It should also be pointed out that some nodes in $T_j[x_j]$ may become suitable pivots after the elimination of x_j or $\{x_j, x_k\}$. For simplicity, we have not taken this into consideration in Algorithm 4.1.

To facilitate the discussion of the tree transformation rules, we introduce a tree manipulation function. Let T be a given tree and x be a node in T which is not the root. Consider the removal of the node x and its incident edges from the tree T so that the children nodes of x (if any) will become the children nodes of the parent of x. This will give rise to another tree, and we shall use remove (x, T) to denote the resulting tree.

The tree transformation rules for Algorithm 4.1 can then be described in terms of the function remove (x, T). Consider the transformation from T_j to T_{j+1}. We have the following three cases.

Case 1. The node x_j is delayed for elimination:

$$T_{j+1} := T_j,$$

Case 2. The node x_j is eliminated as a 1×1 pivot:

$$T_{j+1} := \text{remove } (x_j, T_j),$$

Case 3. $\{x_j, x_k\}$ are eliminated as a 2×2 pivot:

$$T_{j+1} := \text{remove } (x_j, \text{remove } (x_k, T_j)).$$

To illustrate the tree transformation sequence, we use the matrix structure of Fig. 3.1 and its elimination tree of Fig. 3.2. Figure 4.1 displays the sequence of trees resulting from Algorithm 4.1. It corresponds to the following block elimination sequence:

$$x_1, x_3, x_5, \{x_6, x_4\}, x_7, \{x_8, x_2\}.$$

We assume that the nodes x_2 and x_4 have been delayed for elimination until x_8 and x_6, respectively, due to stability consideration on the numerical values.

Note that T_6 = remove (x_5, T_5), and in the tree T_6, the node x_6 becomes the parent of x_4 after the removal of the node x_5 from T_5. On the other hand, T_7 = remove $(x_6,$ remove $(x_4, T_6))$. Since the subtree $T_6[x_6]$ contains only the nodes x_4 and x_6, the tree T_7 is obtained from T_6 by simply the deletion of the entire subtree $T_6[x_6]$.

It is important to point out that a different set of numerical values may induce a different block elimination sequence. Indeed, if the numerical values are given so that no pivoting is necessary (for example, when the matrix is positive definite), it is simple to generate the tree sequence. In such case, in each tree T_j, the subtree $T_j[x_j]$ has the

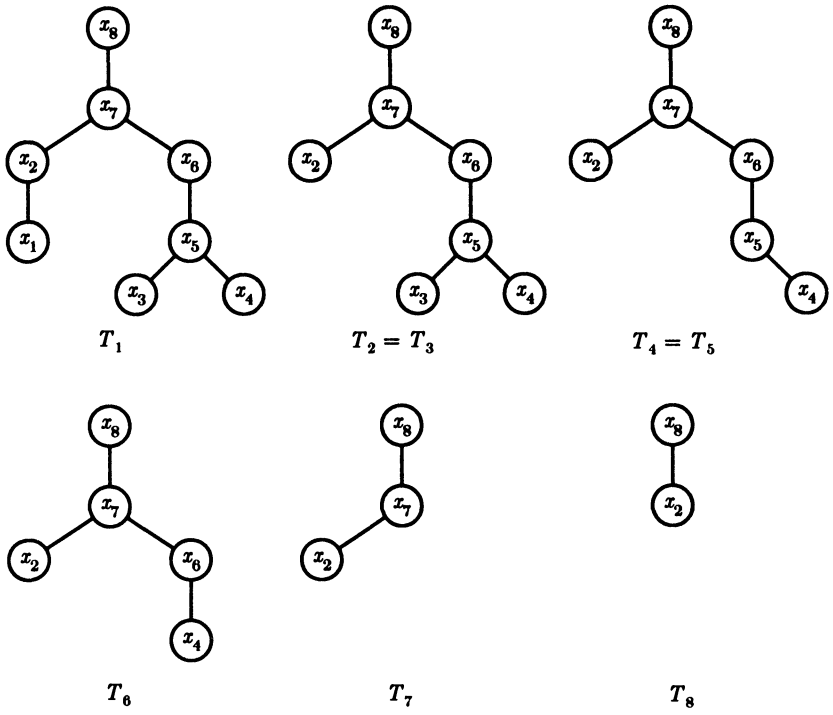


FIG. 4.1. *Tree transformation sequence.*

only node x_j and this node is always a leaf of T_j. Therefore, T_{j+1} can be obtained by simply deleting the leaf node x_j from the tree T_j.

The sequence of tree transformations provides a model to study sparse symmetric indefinite matrix factorization using a mixture of 1×1 and 2×2 block pivots. Implicitly used in the model (or Algorithm 4.1) is the technique of *delayed elimination* as discussed in § 4. How far a node is to be delayed depends on the pivoting strategy and the numerical values. However, irrespective of the strategy, at step j, the subtree $T_j[x_j]$ contains the set of candidates for the next block pivot.

5. Use of the model.

5.1. The multifrontal method by Duff and Reid. The tree model described in § 4 captures the important characteristics of delayed elimination. The tree T_j at step j provides the current structural information necessary for the next elimination. Indeed, the subtree $T_j[x_j]$ rooted at x_j contains the set of desirable candidates for the next pivot, desirable from a structural point of view. It represents the set of nodes in $T[x_j]$, whose elimination has been thus far delayed.

Using the tree model as the basis, we can concentrate on other aspects of sparse indefinite factorization:

(a) Pivoting strategies for numerical stability and factor sparsity,
(b) Algorithms to search for pivots in the subtree $T_j[x_j]$,
(c) Design of data structures to represent the structural and numerical factors,
(d) Subtree representation of T_j,
(e) Efficient numerical sparse factorization,
(f) Forward and backward substitutions.

Different schemes in each of the above categories can be compared with each other using the model as the common basis.

It is appropriate at this point to discuss the relation of the tree model with the multifrontal method by Duff and Reid [7] for sparse indefinite systems. The notion of delayed elimination first appears in their paper [7]. Our work in § 3 considers this notion in the context of elimination trees, and provides some quantitative bounds on possible structural damages due to delayed eliminations. Furthermore, we have incorporated this important idea to give a tree model to study sparse indefinite factorization.

Naturally, there are many ways to implement the tree model depending on the pivoting strategy, pivot searching algorithm, data structures and numerical solution approaches. Indeed, the multifrontal method by Duff and Reid can be treated as one way of implementing the tree model. They have employed the threshold pivoting strategy from [8] for the selection of stable 1×1 and 2×2 pivots. A slightly improved version appears in [12]. Other competitive pivoting schemes exist; one such example is the use of a threshold version of the Bunch–Kaufman pivoting strategy [2], [14].

The main feature in the multifrontal method is the use of full matrices in the course of factorization. Each frontal matrix is stored as a full matrix. This choice of data structure greatly facilitates the search for pivots and adapts extremely well on vector machines. Furthermore, the subtrees $T_j[x_j]$ are represented implicitly in the full submatrix scheme. However, it should be emphasized again that this is only one of many possibilities.

In terms of numerical factorization, the multifrontal method uses an outer product form of factorization. When a row/column is eliminated, its modification to the remaining submatrix is applied. Furthermore, Duff and Reid uses a version of *implicit* (*asymmetric*) *block factorization* whenever the diagonal block is 2×2. In other words, if D is a 2×2 diagonal pivot, and F is the corresponding off-diagonal block, they opt to store D^{-1} and F rather than D and FD^{-1}. This helps to reduce storage as reported in [8].

5.2. New sparse factorization schemes: an example. To illustrate our point that the tree model forms an important basis for different sparse factorization schemes for indefinite matrices, we shall provide one such scheme as an example. It should be stressed that we are not advocating this scheme over other methods, but it serves the purpose of showing the fundamental importance of the model.

A node can be delayed for elimination to be after *any* one of its ancestor nodes in the elimination tree. A simple scheme is to always force the delayed elimination to be after the root x_n of the tree. This actually produces an effective and elegant overall solution method provided that the number of delayed eliminations is relatively small. Algorithm 4.1 to determine block elimination sequence can be reformulated as follows.

ALGORITHM 5.1.

begin
  $T_1 := T(A)$;
  for $j := 1$ **to** n **do**
    begin
      if $x_{j-1} \in T_j[x_j]$ **and** $\{x_{j-1}, x_j\}$ forms a suitable 2×2 pivot
        then eliminate $\{x_{j-1}, x_j\}$ and transform the tree
                $T_{j+1} :=$ remove $(x_j,$ remove $(x_{j-1}, T_j))$
        else
          if x_j is a suitable 1×1 pivot
            then eliminate x_j and transform the tree $T_{j+1} :=$ remove (x_j, T_j)
            else set $T_{j+1} := T_j$ /* delay elimination of x_j */
    end;

eliminate the remaining nodes in T_{n+1}.
end.

Towards the end of the algorithm, the tree T_{n+1} contains all the nodes that have been delayed for elimination. After the numerical elimination of all previous pivots (1×1 $\{x_j\}$ or 2×2 $\{x_{j-1}, x_j\}$), we can treat the matrix remaining to be factored as a dense matrix. In practice, this is almost always the case. The standard routines from LINPACK [6] for factoring indefinite symmetric dense matrices can be used on this submatrix.

A slight improvement in Algorithm 5.1 is in the choice of 2×2 pivots. If $x_c \in T_j[x_j]$ and x_c is a child of x_j in the original elimination tree, then x_c is also a potential candidate for stable 2×2 pivots to go with x_j (without affecting the structure of $\hat{L}$). In the algorithm, c is always taken to be $j-1$.

We can actually consider the scheme from a matrix partitioning point of view. Let A be the given matrix, and P be the permutation matrix corresponding to all the delayed eliminations. Then we can view the permuted matrix as partitioned into:

$$PAP^T = \begin{bmatrix} E & F^T \\ F & C \end{bmatrix},$$

where E is $n-k$ by $n-k$, C is k by k, and k is the number of delayed eliminations in Algorithm 5.1.

This view allows the use of *asymmetric block factorization* [10]:

$$\begin{bmatrix} E & F^T \\ F & C \end{bmatrix} = \begin{bmatrix} E & 0 \\ F & \bar{C} \end{bmatrix} \begin{bmatrix} I & E^{-1}F^T \\ 0 & I \end{bmatrix}.$$

where $\bar{C} = C - FE^{-1}F^T$. Of course, the matrix E itself will be decomposed into its triangular factors, and the pivots will be governed by Algorithm 5.1. On the other hand, $\bar{C}$, which is treated as dense, is factored using a mixture of 1×1 and 2×2 pivots. Note that within $\bar{C}$, we are free to interchange rows/columns without causing structural problems.

In terms of storage, it is important to realize that we need only to store the factors of E and $\bar{C}$ together with the off-diagonal block F of the original matrix. The matrix product $E^{-1}F^T$ is never stored nor computed. An important consequence of this observation is that the compressed data structure obtained by a symbolic factorization of A (see, for example, [10]) is appropriate for the matrix factors of E. Moreover, for those columns corresponding to C in the data structure, their column data storage can be used to keep the associated rows of F. This means the *only* additional data storage required is a full matrix of size k.

For this approach to be successful, it is crucial that the number of delayed eliminations k must be kept as small as possible. Any extra effort to ensure this seems to be worthwhile. Provided k is small, this scheme is quite attractive. Its data structure is similar to those for sparse Cholesky factorization with an additional $k \times k$ full matrix. It also allows an efficient implementation of the numerical factorization and solution phases.

6. Concluding remarks. In sparse Cholesky factorization, the elimination graph model [15], [16] plays a central role. It provides a clear conceptual picture of the elimination process and hence facilitates the development of other important ideas for sparse factorization. Some of the key results include: the improvement on symbolic factorization,

the compressed column data structure by Sherman [18], the significant advance in the implementation of the minimum degree ordering algorithm [10].

The tree model introduced in this paper for sparse symmetric factorization plays a similar role. It captures most of the structural aspects of the elimination process using delayed elimination. This will allow researchers to focus and improve on other algorithmic aspects of factorization based on the model.

There are many ways to implement this tree model, the multifrontal method by Duff and Reid being one of them. The author is currently investigating other practical ways for its efficient implementation. One such scheme is suggested in this paper, which works well if the number of delayed eliminations is small.

REFERENCES

[1] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *Data Structures and Algorithms*, Addison-Wesley, Reading, MA, 1983.

[2] J. R. BUNCH AND L. KAUFMAN, *Some stable methods for calculating inertia and solving symmetric linear equations*, Math. Comp., 31 (1977), pp. 163–179.

[3] J. R. BUNCH, L. KAUFMAN AND B. N. PARLETT, *Decomposition of a symmetric matrix*, Numer. Math., 27 (1976), pp. 95–109.

[4] 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.

[5] A. DAX, *Partial pivoting strategies for symmetric Gaussian elimination*, Math. Programming, 22 (1982), pp. 288–303.

[6] J. J. DONGARRA, J. R. BUNCH, C. B. MOLER AND G. W. STEWART, LINPACK *Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[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] I. S. DUFF, J. K. REID, N. MUNKSGAARD AND H. B. NIELSON, *Direct solution of sets of linear equations whose matrix is large, symmetric and indefinite*, J. Inst. Math. Appl., 23 (1979), pp. 235–250.

[9] R. FLETCHER, *Factorizing symmetric indefinite matrices*, Linear Algebra Appl., 14 (1976), pp. 257–272.

[10] J. A. GEORGE AND J. W. H. LIU, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[11] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, The Johns Hopkins Univ. Press, Baltimore, MD, 1983.

[12] J. W. H. LIU, *On threshold pivoting in the multifrontal method for sparse indefinite systems*, Tech. Report CS-86-06, Dept. of Computer Science, York Univ., North York, Ontario, Canada, 1986.

[13] ———, *A compact row storage scheme for Cholesky factors using elimination trees*, ACM Trans. Math Software, 12 (1986), pp. 127–148.

[14] ———, *A partial pivoting strategy for sparse symmetric matrix decomposition*, ACM Trans. Math Software, 13 (1987), pp. 201–210.

[15] S. PARTER, *The use of linear graphs in Gauss elimination*, SIAM Rev., 3 (1961), pp. 119–130.

[16] D. J. ROSE, *A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations*, in Graph Theory and Computing, R. C. Read, ed., Academic Press, New York, 1972, pp. 183–217.

[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, Yale Univ., New Haven, CT, 1975.

[19] R. E. TARJAN, *Data Structures and Network Algorithms*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 1, January 1988

004

ON THE SPECTRAL DECOMPOSITION OF HERMITIAN MATRICES MODIFIED BY LOW RANK PERTURBATIONS WITH APPLICATIONS*

PETER ARBENZ† AND GENE H. GOLUB‡

Abstract. We consider the problem of computing the eigenvalues and vectors of a matrix $\tilde{H} = H + D$ which is obtained from an indefinite Hermitian low rank modification D of a Hermitian matrix H with known spectral decomposition. It is shown that the eigenvalues of $\tilde{H}$ can easily be located to any desired accuracy by means of the inertia of a Hermitian matrix of small order whose elements depend nonlinearly on the eigenvalue parameter λ. The results are applied to the singular value decomposition of arbitrary modified matrices and to the spectral decomposition of modified unitary and of Hermitian Toeplitz matrices.

For both the singular value decomposition and the unitary eigenvalue problem, divide and conquer algorithms based on rank one modifications are presented.

Key words. Modified eigenvalue problem, Hermitian matrix, Toeplitz matrix, Unitary matrix, Modified singular value decomposition

AMS(MOS) subject classifications. 15A18, 65F15

1. Introduction.

Let $H \in \mathbb{C}^{n\times n}$ be a Hermitian matrix with *known* spectral decomposition

$$H = Q\Lambda Q^{\mathrm{H}} \tag{1}$$

where $Q = [q_1, \cdots, q_n]$ is unitary and $\Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$. $\lambda_1 \geq \cdots \geq \lambda_n$ are the eigenvalues and q_i, $i = 1, \cdots, n$, the corresponding orthonormal eigenvectors of H. We denote the spectrum of H by $\lambda(H) := \{\lambda_1, \cdots, \lambda_n\}$.

Let $D \in \mathbb{C}^{n\times r}$ be an arbitrary Hermitian matrix of small rank $r \leq n$. We consider the problem of finding the eigenvalues and vectors of the matrix

$$\tilde{H} := H + D \tag{2}$$

using the already known spectral decomposition of H. Let $V \in \mathbb{C}^{n\times r}$ be a matrix of maximal rank r with columns spanning $\mathcal{R}(D)$. Then $V(V^{\mathrm{H}}V)^{-1}V^{\mathrm{H}}$ is the orthogonal projection onto $\mathcal{R}(D)$. Therefore, because $\mathcal{N}(D) = \mathcal{R}(D)^{\perp}$ [14, p. 21] we have

$$D = V(V^{\mathrm{H}}V)^{-1}V^{\mathrm{H}}DV(V^{\mathrm{H}}V)^{-1}V^{\mathrm{H}} = V\Delta V^{\mathrm{H}}, \tag{3}$$

where

$$\Delta = (V^{\mathrm{H}}V)^{-1}V^{\mathrm{H}}DV(V^{\mathrm{H}}V)^{-1} \in \mathbb{C}^{r\times r}$$

is a *nonsingular* Hermitian matrix. Thus the problem considered can be reformulated in

$$\tilde{H}x := (H + V\Delta V^{\mathrm{H}})x = \tilde{\lambda}x, \tag{4}$$

*Received by the editors August 10, 1987; accepted for publication (in revised form) November 26, 1987. This work was in part supported by the National Science Foundation under Grant US NSF DCR-8412314 and by the US Army (DAAG 2983-K-0124).

†Visitor, Department of Computer Science, Stanford University, Stanford, California 94305; on leave from BBC Brown, Boveri & Co. Ltd., CH-5401 Baden, Switzerland. Present address: ETH Zürich, Institut für Informatik, CH-8092 Zürich, Switzerland.

‡Department of Computer Science, Stanford University, Stanford, California 94305.

or equivalently according to (1) in

$$(\Lambda + U\Delta U^{\mathrm{H}})y = \tilde{\lambda}y, \quad y = QU,\ x = Qy. \tag{5}$$

If D is *positive semidefinite*, V can be chosen such that Δ is the identity matrix. This special case has been considered by the authors in an earlier paper [2].

In §2 we use a similar analysis to show that the numbers $P(\lambda)$ and $\tilde{P}(\lambda)$ of the eigenvalues of H and $\tilde{H}$ that are $\geq \lambda$, respectively, are related through the equation

$$P(\lambda) + \pi(\Delta^{-1}) = \tilde{P}(\lambda) + \pi(\Delta^{-1} - V^{\mathrm{H}}(\lambda - H)^{-1}V), \quad \lambda \notin \lambda(H), \tag{6}$$

where $\pi(\Delta^{-1})$ and $\pi(\Delta^{-1} - V^{\mathrm{H}}(\lambda - H)^{-1}V)$ are the numbers of *positive* eigenvalues of Δ^{-1} and $\Delta^{-1} - V^{\mathrm{H}}(\lambda - H)^{-1}V$, respectively.

A formula equivalent to (6) has apparently been stated for the first time by Beattie and Fox [4]. A corresponding formula for restricted matrix eigenvalue problems was given by Simpson [23]. Numerical computations using that formula have been performed by Simpson and his collaborators for the frequency analysis of mechanical structures [22], [24].

A slight modification of the matrix $\Delta^{-1} - V^{\mathrm{H}}(\lambda - H)^{-1}V$ permits the extension of statement (6) to the case where λ is an eigenvalue of H. The theory developed in §2 holds for any $r \leq n$. A reasonable application in numerical computations, however, seems to be restricted to an r that is small.

In §3 we apply the theory of §2 to the *modified singular value problem*: Let $A \in \mathbb{C}^{m\times n}, m \geq n$, be a matrix with known singular value decomposition. We consider the problem of computing the singular value decomposition of

$$\tilde{A} := A + X \tag{7}$$

where the perturbation $X \in \mathbb{C}^{m\times n}$ is again assumed to be a matrix of low rank. Because the singular values of A and $\tilde{A}$ are the positive square roots of the eigenvalues of $A^{\mathrm{H}}A$ and of $\tilde{A}^{\mathrm{H}}\tilde{A}$, it is possible to apply the results of §2 setting $H = A^{\mathrm{H}}A$ and $\tilde{H} = \tilde{A}^{\mathrm{H}}\tilde{A}$.

In §4 we show how the *divide and conquer algorithm* that has been proposed by Cuppen for the tridiagonal symmetric eigenvalue problem [8], [10] can be applied on the singular value decomposition of upper bidiagonal matrices. It is surprising that X in (7) can be chosen such that the modification $D = \tilde{A}^{\mathrm{H}}\tilde{A} - A^{\mathrm{H}}A$ has rank one. This is in contrast with the approach made by Jessup and Sorensen [19] which leads to a rank two change.

The results of §2 apply also to the eigenvalue problem of *Toeplitz* matrices and to the eigenvalue problem of *unitary* matrices as will be shown in §§ 5 and 6. In the latter we consider the eigenvalue problem $S^{\mathrm{H}}Ux = \lambda x$, where $U \in \mathbb{C}^{n\times n}$ is a unitary matrix with known spectral decomposition and $S \in \mathbb{C}^{n\times n}$ is a unitary matrix such that $I - S$ has low rank (S may, e.g., be a Householder transformation [17, p. 4]). By means of the Cayley transform [12] the unitary eigenvalue problem can be transformed in a Hermitian one permitting again the application of the theory of §2.

In §7 we discuss some questions that arise when the derived results are to be applied numerically.

2. Locating the eigenvalues of $\tilde{H} = H + V\Delta V^{\mathrm{H}}$. In [1], [2] the eigenvalue problem (5) with $\Delta = I$ has been investigated. In this section we generalize the analysis developed there to the more general case where Δ is an *arbitrary nonsingular* matrix. We first show the basic lemma.

LEMMA 2.1. *Let $\lambda \in \mathbb{R}$ be arbitrary but fixed and let $\mu = \mu(\lambda) \geq 0$ be the multiplicity of λ as eigenvalue of H. Let $W \in \mathbb{C}^{n\times\mu}$ be a matrix the columns of which form an orthonormal basis of the eigenspace $\mathcal{N}(\lambda - H)$ corresponding to λ. Then the matrices*

$$B := \begin{bmatrix} \lambda - \tilde{H} & 0 & 0 \\ 0 & \Delta^{-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} \in \mathbb{C}^{(n+r+\mu)\times(n+r+\mu)} \tag{8}$$

and

$$C := \begin{bmatrix} \lambda - H & 0 & 0 \\ 0 & \Delta^{-1} - V^{\mathrm{H}}(\lambda - H)^{+}V & V^{\mathrm{H}}W \\ 0 & W^{\mathrm{H}}V & 0 \end{bmatrix} \in \mathbb{C}^{(n+r+\mu)\times(n+r+\mu)} \tag{9}$$

are congruent. Here $(\lambda - H)^{+}$ is the Moore-Penrose pseudoinverse of $(\lambda - H)$.

Proof. A simple computation shows that

$$M^{\mathrm{H}}BM = C \tag{10}$$

with

$$\begin{aligned} M := & \begin{bmatrix} I - WW^{\mathrm{H}} & 0 & W \\ 0 & I & 0 \\ W^{\mathrm{H}} & 0 & 0 \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ \Delta V^{\mathrm{H}}(I - WW^{\mathrm{H}}) & I & 0 \\ 0 & 0 & I \end{bmatrix} \\ & \cdot \begin{bmatrix} I & 0 & 0 \\ 0 & I & \Delta V^{\mathrm{H}}W \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & -(\lambda - H)^{+}V & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \\ = & \begin{bmatrix} I - WW^{\mathrm{H}} & -(\lambda - H)^{+}V & W \\ \Delta V^{\mathrm{H}}(I - WW^{\mathrm{H}}) & I - \Delta V^{\mathrm{H}}(\lambda - H)^{+}V & \Delta V^{\mathrm{H}}W \\ W^{\mathrm{H}} & 0 & 0 \end{bmatrix} \end{aligned} \tag{11}$$

Recall that $(\lambda - H)^{+}(\lambda - H) = I - WW^{\mathrm{H}}$. □

If λ is not an eigenvalue of H then Lemma 2.1 reduces to Corollary 2.2.

COROLLARY 2.2. *If $\lambda \notin \lambda(H)$, then the matrices*

$$B := \begin{bmatrix} \lambda - \tilde{H} & 0 \\ 0 & \Delta^{-1} \end{bmatrix} \in \mathbb{C}^{(n+r)\times(n+r)} \tag{12}$$

and

$$C := \begin{bmatrix} \lambda - H & 0 \\ 0 & \Delta^{-1} - V^{\mathrm{H}}(\lambda - H)^{-1}V \end{bmatrix} \in \mathbb{C}^{(n+r)\times(n+r)} \tag{13}$$

are congruent.

Proof. Equation (10) holds with

$$M := \begin{bmatrix} I & -(\lambda - H)^{-1}V \\ \Delta V^{\mathrm{H}} & I - \Delta V^{\mathrm{H}}(\lambda - H)^{-1}V \end{bmatrix}. \tag{14}$$

□

Since $\det M = 1$, (10) yields

$$\frac{\prod_{j=1}^{n}(\lambda-\tilde{\lambda}_j)}{\prod_{j=1}^{n}(\lambda-\lambda_j)} = \det(\Delta)\det(\Delta^{-1}-V^{\mathrm{H}}(\lambda-H)^{-1}V), \quad \lambda \notin \lambda(H). \tag{15}$$

Thus, the eigenvalues of $\tilde{H}$ can in principle be obtained from those of H by an investigation of the zeros and poles of the function on the right-hand side of (15). This function is a generalization of the *Weinstein–Aronszajn determinant* known from the methods of intermediate problems [2], [15], [27].

Let $P(\lambda)$ and $\tilde{P}(\lambda)$ be the number of eigenvalues of H and $\tilde{H}$, respectively, that are $\geq \lambda$. We denote by $(\nu(A), \varsigma(A), \pi(A))$ the *inertia* [14], [20], i.e., the number of negative, zero, and positive eigenvalues of a Hermitian matrix A. Then we have Theorem 2.3.

THEOREM 2.3. *Let $\lambda \in \mathbb{R}$ be arbitrary but fixed and let μ be the multiplicity of λ as eigenvalue of H. Let $W \in \mathbb{C}^{n\times\mu}, W^{\mathrm{H}}W = I_\mu$, be such that $\mathcal{R}(W) = \mathcal{N}(\lambda - H)$. Then the equality*

$$\tilde{P}(\lambda) + \pi(Z(\lambda)) = P(\lambda) + \pi(\Delta) \tag{16}$$

holds with

$$Z(\lambda) = \begin{bmatrix} \Delta^{-1} - V^{\mathrm{H}}(\lambda - H)^{+}V & V^{\mathrm{H}}W \\ W^{\mathrm{H}}V & 0 \end{bmatrix} \in \mathbb{C}^{(r+\mu)\times(r+\mu)}. \tag{17}$$

Furthermore, the mapping

$$\begin{aligned} \alpha:\ \mathcal{N}(Z(\lambda)) &\longrightarrow \mathcal{N}(\lambda - \tilde{H}) \\ \begin{bmatrix} y \\ z \end{bmatrix} &\longmapsto x := (\lambda - H)^{+}Vy + Wz \end{aligned} \tag{18}$$

is bijective.

Proof. By Sylvester's law B and C have the same inertia [20]. The number of nonpositive eigenvalues of B is $\tilde{P}(\lambda) + r - \pi(\Delta^{-1}) + \mu$. This number equals $P(\lambda) + r + \mu - \pi(Z(\lambda))$, the number of nonpositive eigenvalues of C. As $\pi(\Delta^{-1}) = \pi(\Delta)$, this proves (16).

M in (11) bijectively maps

$$\begin{aligned} \mathcal{N}(C) = &\left\{ \begin{bmatrix} x \\ 0 \end{bmatrix} \in \mathbb{C}^{n+r+\mu} \middle| x \in \mathcal{N}(\lambda - H) \subset \mathbb{C}^n \right\} \\ &\oplus \left\{ \begin{bmatrix} 0 \\ y \\ z \end{bmatrix} \in \mathbb{C}^{n+r+\mu} \middle| \begin{bmatrix} y \\ z \end{bmatrix} \in \mathcal{N}(Z) \subset \mathbb{C}^{r+\mu} \right\} \end{aligned} \tag{19}$$

onto

$$\mathcal{N}(B) = \left\{ \begin{bmatrix} x \\ 0 \\ 0 \end{bmatrix} \in \mathbb{C}^{n+r+\mu} \middle| x \in \mathcal{N}(\lambda - \tilde{H}) \right\} \oplus \left\{ \begin{bmatrix} 0 \\ 0 \\ z \end{bmatrix} \in \mathbb{C}^{n+r+\mu} \middle| z \in \mathbb{C}^{\mu} \right\}. \tag{20}$$

From (11) it is seen that the first summand in (19) is mapped bijectively onto the second summand in (20) by

$$x \longmapsto z = W^{\mathrm{H}}x, \tag{21}$$

while the second summand of $\mathcal{N}(C)$ is bijectively mapped onto the first of $\mathcal{N}(B)$ by α. This completes the proof. □

If λ is *not* an eigenvalue of H then Theorem 2.3 reduces to Theorem 2.4.

THEOREM 2.4. *Let $\lambda \notin \lambda(H)$, i.e., $\mu(\lambda) = 0$, and*

$$Z(\lambda) = \Delta^{-1} - V^{\mathrm{H}}(\lambda - H)^{-1}V. \tag{22}$$

Then (16) holds and the mapping

$$\begin{aligned} \alpha:\ \mathcal{N}(Z(\lambda)) &\longrightarrow \mathcal{N}(\lambda - \tilde{H}) \\ y &\longmapsto x := (\lambda - H)^{-1}Vy \end{aligned} \tag{23}$$

is bijective.

Proof. Theorem 2.4 follows similarly from Corollary 2.2 as Theorem 2.3 from Lemma 2.1. □

Remarks. (i) We may write (16) in the form

$$\tilde{\lambda}_{\tilde{P}(\lambda)} \geq \lambda \tilde{\lambda}_{\tilde{P}(\lambda)+1}, \tag{24}$$

where $\tilde{P}(\lambda) = P(\lambda) + \pi(\Delta) - \pi(Z(\lambda))$, or, since $\varsigma(Z(\lambda))$ is the multiplicity of λ as eigenvalue of $\tilde{H}$, in the form

$$\tilde{\lambda}_{\tilde{P}(\lambda)-\varsigma(Z(\lambda))} > \lambda \geq \tilde{\lambda}_{\tilde{P}(\lambda)-\varsigma(Z(\lambda))+1}. \tag{25}$$

These inequalities render it possible to compute any eigenvalue of $\tilde{H}$ to any desired accuracy by a bisection algorithm [14].

(ii) If we define $N(\lambda)$ and $\tilde{N}(\lambda)$ to be the number of eigenvalues of H and $\tilde{H}$ that are $\leq \lambda$ then the inequality

$$\tilde{N}(\lambda) + \nu(Z(\lambda)) = N(\lambda) + \nu(\Delta) \tag{26}$$

holds, the proof being similar to the one of (16).

Note that, using the spectral decomposition of H, $Z(\lambda)$ in (22) can be written as

$$Z(\lambda) = \Delta^{-1} - U^{\mathrm{H}}(\lambda - \Lambda)^{-1}U. \tag{27}$$

Since $\lambda - \Lambda$ is diagonal the computation of $Z(\lambda)$ is cheap. If r is sufficiently small it may thus be advantageous to compute the eigenvalues and vectors by means of (26) instead of forming the whole matrix $\tilde{H}$ and apply one of the classical algorithms such as the QR-algorithm.

The following Theorem gives useful a priori inclusions for the eigenvalues of $\tilde{H}$.

THEOREM 2.5. *Let $r_+ = \pi(\Delta)$ and $r_- = \nu(\Delta) = r - r_+$. Setting $\lambda_j = +\infty$ for $j < 1$ and $\lambda_j = -\infty$ for $j > n$, the inequalities*

$$\lambda_{j-r_+} \geq \tilde{\lambda}_j \geq \lambda_{j+r_-}, \quad j = 1, \cdots, n \tag{28}$$

are valid.

Proof. Evaluating (16) at the point $\tilde{\lambda}_j$ we obtain

$$\begin{aligned} P(\tilde{\lambda}_j) &= \tilde{P}(\tilde{\lambda}_j) - \pi(\Delta) + \pi(Z(\tilde{\lambda}_j)) \\ &\geq \tilde{P}(\tilde{\lambda}_j) - r_+ \geq j - r_+ \end{aligned}$$

which proves the left-hand-side inequality. Similarly, by evaluating (16) at λ_j one obtains

$$\begin{aligned} \tilde{P}(\lambda_j) &= P(\lambda_j) + \pi(\Delta) - \pi(Z(\lambda_j)) \\ &\geq j + r - r_- - r + \nu(Z(\lambda_j)) + \varsigma(Z(\lambda_j)) \\ &\geq j - r_-, \end{aligned}$$

from which the right-hand-side inequality follows. □

Remark. Inequalities (28) are well known in the case $r_+ = r$ ($r_- = 0$) [26], [27]. They are usually proved by means of the Courant–Weyl principle. The use of this principle is now hidden in the proof of Sylvester's law on inertia [14], [20] which was used in the proof of Theorem 2.3.

3. On the modified singular value decomposition. Let now $A \in \mathbb{C}^{m\times n}$, $m \geq n$, be a matrix with *known* singular value decomposition

$$A = F\Sigma G^{\mathrm{H}}, \quad \text{where } F \in \mathbb{C}^{m\times m} \text{ and } G \in \mathbb{C}^{n\times n} \text{ are unitary matrices} \tag{29}$$

and

$$\Sigma = \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix} \in \mathbb{C}^{m\times n}, \qquad \Sigma_1 = \operatorname{diag}(\sigma_1, \cdots, \sigma_n), \tag{30}$$

contains the singular values $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ of A in its diagonal. We consider the problem of computing the singular value decomposition

$$\tilde{A} = \tilde{F}\tilde{\Sigma}\tilde{G}^{\mathrm{H}} \tag{31}$$

of the matrix $\tilde{A} := A + X$. Here it is assumed that X has low rank, say p. This implies that X can be represented in the form

$$X = X_m X_n^{\mathrm{H}}, \tag{32}$$

where both $X_m \in \mathbb{C}^{m\times p}$ and $X_n \in \mathbb{C}^{n\times p}$ have maximal rank p.

As is well known, the singular values of A and $\tilde{A}$ are the positive square roots of the eigenvalues of the positive semidefinite matrices $H := A^{\mathrm{H}}A = G\Sigma^2 G^{\mathrm{H}}$ and

$$\tilde{H} := \tilde{A}^{\mathrm{H}}\tilde{A} = (A+X)^{\mathrm{H}}(A+X) = H + D \in \mathbb{C}^{n\times n}, \tag{33}$$

respectively, where

$$\begin{aligned} D &:= A^{\mathrm{H}}X + X^{\mathrm{H}}A + X^{\mathrm{H}}X \\ &= A^{\mathrm{H}}X_m X_n^{\mathrm{H}} + X_n X_m^{\mathrm{H}} A + X_n X_m^{\mathrm{H}} X_m X_n^{\mathrm{H}} \end{aligned} \tag{34}$$

is a Hermitian and in general indefinite matrix. From (34) it is seen that its range $\mathcal{R}(D)$ is spanned by the columns of X and $A^{\mathrm{H}}X$. Recalling that $\operatorname{rank} X = p$ we get $p \leq r := \dim \mathcal{R}(D) \leq 2p$.

As shown in §1, D can be represented in the form

$$D = V\Delta V^{\mathrm{H}}, \quad V \in \mathbb{C}^{n\times r}, \quad \Delta \in \mathbb{C}^{r\times r}, \tag{35}$$

where Δ is diagonal. Then Theorem 2.3 is easily rewritten for the present case.

THEOREM 3.1. *Let $\lambda \in \mathbb{R}$ be arbitrary and let μ be the multiplicity of $\sigma := \operatorname{sign}(\lambda)\sqrt{|\lambda|}$ as singular value of A. Let $P(\lambda)$ and $\tilde{P}(\lambda)$ denote the numbers of singular values of A and $\tilde{A}$, respectively, that are $\geq \sigma$ and let the columns of $W \in \mathbb{C}^{n\times\mu}$, form an orthonormal basis of $\mathcal{N}(\lambda - A^{\mathrm{H}}A)$. Then the equality*

$$\tilde{P}(\lambda) + \pi(Z(\lambda)) = P(\lambda) + \pi(\Delta) \tag{36}$$

holds with

$$Z(\lambda) := \begin{bmatrix} \Delta^{-1} - V^{\mathrm{H}}(\lambda - A^{\mathrm{H}}A)^{+}V & V^{\mathrm{H}}W \\ W^{\mathrm{H}}V & 0 \end{bmatrix}. \tag{37}$$

Furthermore, the mapping

$$\begin{aligned} \alpha:\ \mathcal{N}(Z(\lambda)) &\longrightarrow \mathcal{N}(\lambda - \tilde{A}^{\mathrm{H}}\tilde{A}) \\ \begin{bmatrix} y \\ z \end{bmatrix} &\longmapsto x := (\lambda - A^{\mathrm{H}}A)^{+}Vy + Wz \end{aligned} \tag{38}$$

is bijective.

Note that the matrix W can be formed with the columns of G corresponding to the singular value σ.

For σ not a singular value of A we get with (29)

$$\begin{aligned} Z(\lambda) &= \Delta^{-1} - V^{\mathrm{H}}(\lambda - A^{\mathrm{H}}A)^{-1}V, \\ &= \Delta^{-1} - (G^{\mathrm{H}}V)^{\mathrm{H}}(\lambda - \Sigma^2)^{-1}(G^{\mathrm{H}}V), \end{aligned} \tag{39}$$

with $\lambda = \operatorname{sign}(\sigma)\sigma^2$, a substitution that makes a repeated computation of $Z(\lambda)$ much cheaper.

The mapping α gives the columns of $\tilde{G}$ in (31) corresponding to the singular value σ. The corresponding columns of $\tilde{F}$ could be obtained in a similar way if one works with $\tilde{A}\tilde{A}^{\mathrm{H}}$ instead of $\tilde{A}^{\mathrm{H}}\tilde{A}$. A better way to get $\tilde{F}$ is via the QR–decomposition with column pivoting of the matrix $\tilde{A}\tilde{G}$ [14, p. 289]:

$$\tilde{Q}\tilde{R} = \tilde{A}\tilde{G}\tilde{\Pi}. \tag{40}$$

Here $\tilde{\Pi}$ is a permutation matrix. It is easy to see that $\tilde{R}$ has orthogonal columns. Indeed

$$\tilde{R}^{\mathrm{H}}\tilde{R} = \tilde{\Pi}^{\mathrm{T}}\tilde{G}^{\mathrm{H}}\tilde{A}^{\mathrm{H}}\tilde{Q}\tilde{Q}^{\mathrm{H}}\tilde{A}\tilde{G}\tilde{\Pi} = \tilde{\Pi}^{\mathrm{T}}\tilde{\Sigma}^2\tilde{\Pi} = (\tilde{\Pi}^{\mathrm{T}}\tilde{\Sigma}\tilde{\Pi})^2$$

with $\tilde{\Sigma} = \operatorname{diag}(\tilde{\sigma}_1, \cdots, \tilde{\sigma}_n)$. Without loss of generality we can assume that the diagonal elements of $\tilde{R}$ are nonnegative. Therefore $\tilde{R} = \Pi^{\mathrm{T}}\tilde{\Sigma}\tilde{\Pi}$ and with (40) we obtain

$$\tilde{A} = (\tilde{Q}\tilde{\Pi}^{\mathrm{T}})\tilde{\Sigma}\tilde{G}, \tag{41}$$

the desired singular value decomposition of $\tilde{A}$.

We now apply Theorem 2.5 to obtain an a priori inequality for singular values. To do that we estimate the number of positive and negative eigenvalues of D. To that

end we choose a $V_0 \in \mathbb{C}^{n\times r}$ of the form $V_0 = [V_1, V_2]$, where $V_1 \in \mathbb{C}^{n\times p}$, $V_1^{\mathrm{H}}V_1 = I_p$, spans $\mathcal{R}(X_n)$ and $V_2 \in \mathbb{C}^{n\times(r-p)}$, $V_2^{\mathrm{H}}V_2 = I_{r-p}$, spans $\mathcal{R}(A^{\mathrm{H}}X_m) \cap \mathcal{R}(X_n)^{\perp}$. Then evidentially $V_2^{\mathrm{H}}V_1 = 0$ and consequently $V_2^{\mathrm{H}}X_n = 0$. Any $u \in \mathcal{R}(D)$ can be represented by

$$u = [V_1, V_2]\begin{bmatrix} x \\ y \end{bmatrix}, \quad x \in \mathbb{C}^p, y \in \mathbb{C}^{r-p} \tag{42}$$

Thus, to any eigenpair $(\bar{\lambda}, \bar{u})$ of D there corresponds a vector $\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} = V_0^{+}\bar{u} = V_0^{\mathrm{H}}\bar{u}$ satisfying

$$V_0^{\mathrm{H}}DV_0\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} = \begin{bmatrix} B & C \\ C^{\mathrm{H}} & 0 \end{bmatrix}\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} = \bar{\lambda}\begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \tag{43}$$

where

$$B = V_1^{\mathrm{H}}A^{\mathrm{H}}X_mX_n^{\mathrm{H}}V_1 + V_1^{\mathrm{H}}X_nX_m^{\mathrm{H}}AV_1 + V_1^{\mathrm{H}}X_nX_m^{\mathrm{H}}X_mX_n^{\mathrm{H}}V_1$$

and

$$C = V_1^{\mathrm{H}}X_nX_m^{\mathrm{H}}AV_2.$$

C clearly has maximal rank $r - p$. Let $\bar{C} \in \mathbb{C}^{p\times p}$ be a nonsingular matrix such that

$$\bar{C}C = \begin{bmatrix} I_{r-p} \\ 0 \end{bmatrix}.$$

Then, since the inertia of a matrix is invariant under congruence transforms, the matrix in (43),

$$\begin{bmatrix} \bar{C} & 0 \\ 0 & I_{r-p} \end{bmatrix}\begin{bmatrix} B & C \\ C^{\mathrm{H}} & 0 \end{bmatrix}\begin{bmatrix} \bar{C}^{\mathrm{H}} & 0 \\ 0 & I_{r-p} \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} & I_{r-p} \\ B_{12}^{\mathrm{H}} & B_{22} & 0 \\ I_{r-p} & 0 & 0 \end{bmatrix} \tag{44}$$

and

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & -B_{12}^{\mathrm{H}} \\ 0 & 0 & I \end{bmatrix}\begin{bmatrix} B_{11} & B_{12} & I_{r-p} \\ B_{12}^{\mathrm{H}} & B_{22} & 0 \\ I_{r-p} & 0 & 0 \end{bmatrix}\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -B_{12} & I \end{bmatrix} = \begin{bmatrix} B_{11} & 0 & I_{r-p} \\ 0 & B_{22} & 0 \\ I_{r-p} & 0 & 0 \end{bmatrix} \tag{45}$$

have the same number of positive and negative eigenvalues. We denote these numbers by r_+ and r_-. Since the mentioned matrices are regular we have $r = r_+ + r_-$. Furthermore, by (45) we see that $r - p \le r_- \le p$ and $r - p \le r_+ \le p$ (cf. [2]). With Theorem 2.5 we thus obtain the a priori inequalities

$$\sigma_{j-p} \ge \sigma_{j-r_+} \ge \tilde{\sigma}_j \ge \sigma_{j+r_-} \ge \sigma_{j+p}. \tag{46}$$

The inclusion $\sigma_{j-p} \ge \tilde{\sigma}_j \ge \sigma_{j+p}$ is well known [25].

4. A special case: A divide and conquer algorithm for the singular value decomposition of bidiagonal matrices. In 1981 Cuppen [8] proposed a divide and conquer algorithm for the solution of the real symmetric tridiagonal eigenvalue problem. The method has been refined and successfully implemented for vector

and parallel computers by Dongarra and Sorensen [10]. After the discussion in the previous section it is of interest to try to generalize this algorithm to the computation of the singular value decomposition of bidiagonal matrices. This has been done by Jessup and Sorensen [19]. In this section we reconsider the problem, treating it in a different way in order for D in (34) to become a matrix of rank one.

Let

$$\tilde{A} = \begin{bmatrix} \delta_1 & \nu_2 & & \\ & \ddots & \ddots & \\ & & \ddots & \nu_n \\ & & & \delta_n \end{bmatrix} \in \mathbb{C}^{n\times n}, \quad \nu_i \neq 0, \tag{47}$$

be a bidiagonal matrix the singular value decomposition of which is to be found. Without loss of generality $\tilde{A}$ can be assumed to be *square*. It may, for example, have been obtained from any $m \times n$–matrix by a finite sequence of Householder transformations [14, p. 170].

We decompose $\tilde{A}$ in the form $\tilde{A} =: A+X = A+e_k v^{\mathrm{H}}$ where $v^{\mathrm{H}} = \delta_k e_k^{\mathrm{T}} + \nu_{k+1} e_{k+1}^{\mathrm{T}}$, e_j denotes the j'th unit vector,

$$A = \begin{bmatrix} \delta_1 & \nu_2 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & \delta_{k-1} & \nu_k & & & & \\ & & & 0 & 0 & & & \\ & & & & \delta_{k+1} & \nu_{k+2} & & \\ & & & & & \ddots & \ddots & \\ & & & & & & \ddots & \nu_n \\ & & & & & & & \delta_n \end{bmatrix},$$

and

$$X = \begin{bmatrix} 0 & 0 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & 0 & 0 & & & & \\ & & & \delta_k & \nu_{k+1} & & & \\ & & & & 0 & 0 & & \\ & & & & & \ddots & \ddots & \\ & & & & & & 0 & 0 \\ & & & & & & & 0 \end{bmatrix}.$$

The matrix $A^{\mathrm{H}}A$ consists of two tridiagonal blocks,

$$A^{\mathrm{H}}A = \begin{bmatrix} (A^{\mathrm{H}}A)_1 & 0 \\ 0 & (A^{\mathrm{H}}A)_2 \end{bmatrix}, \tag{48}$$

with $(A^{\mathrm{H}}A)_1 \in \mathbb{C}^{k\times k}$ and $(A^{\mathrm{H}}A)_2 \in \mathbb{C}^{(n-k)\times(n-k)}$, the first of which is always singular. Because $A^{\mathrm{H}}X = 0$, the matrix D in (33)–(34) becomes

$$D = X^{\mathrm{H}}X = vv^{\mathrm{H}} = [e_k, e_{k+1}] \begin{bmatrix} |\delta_k|^2 & \bar{\delta}_k \nu_{k+1} \\ \delta_k \bar{\nu}_{k+1} & |\nu_{k+1}|^2 \end{bmatrix} [e_k, e_{k+1}]^{\mathrm{H}} \tag{49}$$

which is equivalent with (35) if one sets $\Delta = 1$.

Remark. Jessup and Sorensen chose $X = \nu_{k+1} e_k e_{k+1}^{\mathrm{T}}$, whence a rank two modification D results.

The matrix

$$G = \begin{bmatrix} G_1 & 0 \\ 0 & G_2 \end{bmatrix} \tag{50}$$

in (29) contains the normalized eigenvectors of $A^{\mathrm{H}}A$. It is blockstructured in the same way as $A^{\mathrm{H}}A$ is. Therefore $Z(\lambda)$ in (39) becomes

$$\begin{aligned} Z(\lambda) &= 1 - v^{\mathrm{H}}(\lambda - A^{\mathrm{H}}A)^{-1}v = 1 - (G^{\mathrm{H}}v)^{\mathrm{H}}(\lambda - \Sigma^2)^{-1}G^{\mathrm{H}}v \\ &= 1 - |\delta_k|^2 \sum_{j=1}^{k} \frac{|g_{kj}|^2}{\lambda - \sigma_j^2} - |\nu_{k+1}|^2 \sum_{j=k+1}^{n} \frac{|g_{k+1,j}|^2}{\lambda - \sigma_j^2}, \quad \lambda \notin \lambda(A^{\mathrm{H}}A). \end{aligned} \tag{51}$$

The treatment of the function $Z(\lambda)$ is crucial for the success of a divide and conquer algorithm. As already mentioned after (15), $Z(\lambda)$ contains all the information needed to determine $\lambda(\tilde{A}^{\mathrm{H}}\tilde{A})$. Although it is easy by Theorem 3.1 to determine the multiplicity of every $\lambda_j \in \lambda(A^{\mathrm{H}}A)$ as eigenvalue of $\tilde{A}^{\mathrm{H}}\tilde{A}$ it is preferable in the rank one case to perform the following *deflation process*:

Observe that, since [10]

$$\begin{aligned} \|(\tilde{A}^{\mathrm{H}}\tilde{A} - \sigma_j^2)Ge_j\| &= \|G(\Sigma^2 - \sigma_j^2 + (G^{\mathrm{H}}v)(G^{\mathrm{H}}v)^{\mathrm{H}})e_j\| \\ &= |(G^{\mathrm{H}}v)^{\mathrm{H}}e_j| = |v^{\mathrm{H}}Ge_j| = |(G^{\mathrm{H}}v)_j|, \end{aligned} \tag{52}$$

σ_j^2 is a good approximation for an eigenvalue of $\tilde{A}^{\mathrm{H}}\tilde{A}$ if $|(G^{\mathrm{H}}v)_j|$ is small. This is indeed very often the case as has been observed in the tridiagonal case by Cuppen [8] and Dongarra and Sorensen [10] as well. Therefore those summands in (51) with sufficiently small coefficients can be neglected, or equivalently, the rows and columns corresponding to small $|(G^{\mathrm{H}}v)_j|$ can be eliminated from $A^{\mathrm{H}}A$. (For details see [10].)

A special case of the above situation occurs if $A^{\mathrm{H}}A$ has *multiple* eigenvalues. Then the basis of the corresponding eigenspace can be chosen such that at most one of the eigenvectors is *not* orthogonal to v. By consequence the corresponding components of $G^{\mathrm{H}}v$ vanish and can be eliminated in the deflation process.

Finally one is lead to matrices $A^{\mathrm{H}}A$ and $\tilde{A}^{\mathrm{H}}\tilde{A}$ (we do not change notation) which both have only *simple* eigenvalues. By (15) and the form (51) of $Z(\lambda)$ it is clear that $\lambda(A^{\mathrm{H}}A) \cap \lambda(\tilde{A}^{\mathrm{H}}\tilde{A}) = \emptyset$. Because $Z'(\lambda) > 0$ for all $\lambda \notin \lambda(A^{\mathrm{H}}A)$, $Z(\lambda)$ has a single simple zero in each open interval $(\sigma_j^2, \sigma_{j+1}^2), 1 \leq j < n$, and (σ_n^2, ∞). Bunch et al. [7] have developed a quadratically convergent zerofinder for the determination of the roots of $Z(\lambda) = 0$ based on a rational approximation of $Z(\lambda)$. This zerofinder proved to be very efficient in the symmetric tridiagonal case [10].

If a zero $\tilde{\lambda}$ of $Z(\lambda)$ (i.e., the square of a singular value $\tilde{\sigma}$ of $\tilde{A}$) is found, a corresponding *right* singular vector $\tilde{g}$ is – according to (38) – given by

$$\tilde{g} = (\tilde{\lambda} - A^{\mathrm{H}}A)^{-1}v = G(\tilde{\lambda} - \Sigma^2)^{-1}G^{\mathrm{H}}v. \tag{53}$$

In the absence of round off, a corresponding *left* singular vector is obtained by

$$\tilde{f} = \frac{1}{\tilde{\sigma}}\tilde{A}\tilde{g}. \tag{54}$$

$\tilde{f}$ is normalized if $\tilde{g}$ is so. Note that $\tilde{\sigma}$ does not vanish, since possible singular values zero of $\tilde{A}$ have been ruled out in the deflation process!

To obtain the singular value decomposition of the original matrix $\tilde{A}$, the newly computed singular values and vectors have to be combined with the deflated ones.

Equation (54) may cause inaccurate results if $\tilde{\sigma}$ is very small, possessing a large relative error. To control the accuracy of the computed results we propose to use (54) together with its dual formula

$$\tilde{g} = \frac{1}{\tilde{\sigma}} \tilde{A}^{\mathrm{H}} \tilde{f} \tag{55}$$

in the following *Lanczos process*. Starting with $j = 1$ one defines [13]

$$z_j := \tilde{A}\hat{g}_j, \quad \hat{\sigma}_j := \|z_j\|_2, \quad \hat{f}_j := z_j/\hat{\sigma}_j, \tag{56}$$

where $\hat{g}_j$, $\hat{f}_j$, and $\hat{\sigma}_j$ are approximations to $\tilde{g}_j$, $\tilde{f}_j$, and $\tilde{\sigma}_j$. We consider them to be accurate if $\hat{\beta}_j$ defined by

$$w_j := \tilde{A}^{\mathrm{H}} z_j - \hat{\sigma}_j \hat{g}_j, \quad \hat{\beta}_j := \|w_j\|_2, \tag{57}$$

is sufficiently small, i.e., if its size is of the order of magnitude of the machine precision [14, p. 33]. If this is not the case, we define

$$\hat{g}_{j+1} := w_j/\hat{\beta}_j \tag{58}$$

and perform the Lanczos steps (56)–(57) again starting with $\hat{g}_{j+1}$. If we now assume that $\hat{\sigma}_1, \cdots, \hat{\sigma}_{j-1}$, $\hat{g}_1, \cdots, \hat{g}_{j-1}$, and $\hat{f}_1, \cdots, \hat{f}_{j-1}$ are accurate to machine precision, we can expect the error of $\hat{g}_j$ to lie essentially in the span of $\tilde{g}_{j+1}, \cdots, \tilde{g}_n$. Therefore, by well-known properties of the Lanczos algorithm [14, p. 323], the larger singular value of the matrix

$$\begin{bmatrix} \hat{\sigma}_j & \hat{\beta}_j \\ 0 & \hat{\sigma}_{j+1} \end{bmatrix} \tag{59}$$

is a better approximation to $\tilde{\sigma}_j$ than $\hat{\sigma}_j$. Improved approximations to $\tilde{g}_j$ and $\tilde{f}_j$ are given by $\gamma_1\hat{g}_j + \gamma_2\hat{g}_{j+1}$ and $\varphi_1\hat{f}_j + \varphi_2\hat{f}_{j+1}$, respectively, where $\binom{\gamma_1}{\gamma_2}$ and $\binom{\varphi_1}{\varphi_2}$ are the right and left singular vectors corresponding to the larger singular value of the matrix in (59). If $\hat{\beta}_{j=1} = \|\tilde{A}^{\mathrm{H}}\hat{f}_{j+1} - \sigma_{j+1}\hat{g}_{j+1}\|$ is not small either, the Lanczos process (56)–(58) may be continued. We believe, however, that one or two steps will in general suffice to obtain high accuracy.

5. On the banded Toeplitz eigenvalue problem. Let us consider the eigenvalue problem

$$\tilde{H}x = \tilde{\lambda}x \tag{60}$$

where $\tilde{H}$ is a banded Hermitian Toeplitz matrix

$$\tilde{H} = \begin{bmatrix} t_0 & \cdots & t_p & & \\ \vdots & \ddots & & \ddots & \\ t_p & & \ddots & & t_p \\ & \ddots & & \ddots & \vdots \\ & & t_p & \cdots & t_0 \end{bmatrix}$$

of order n. (We assume $2p < n$.) Jain [18] proposed to solve the linear equation $\tilde{H}x = y$ by decomposing $\tilde{H}$ in the form $\tilde{H} = H - D$ with

$$H = \begin{bmatrix} t_1 & \cdots & t_p & & & t_p & \cdots & t_1 \\ \vdots & \ddots & & \ddots & & & \ddots & \vdots \\ t_p & & \ddots & & \ddots & & & t_p \\ & \ddots & & \ddots & & \ddots & & \\ & & \ddots & & \ddots & & \ddots & \\ t_p & & & \ddots & & \ddots & & t_p \\ \vdots & \ddots & & & \ddots & & \ddots & \vdots \\ t_1 & \cdots & t_p & & & t_p & \cdots & t_0 \end{bmatrix}$$

and

$$D = \begin{bmatrix} 0 & \cdots & 0 & & & t_p & \cdots & t_1 \\ \vdots & \ddots & & \ddots & & & \ddots & \vdots \\ 0 & & \ddots & & \ddots & & & t_p \\ & \ddots & & \ddots & & \ddots & & \\ & & \ddots & & \ddots & & \ddots & \\ t_p & & & \ddots & & \ddots & & 0 \\ \vdots & \ddots & & & \ddots & & \ddots & \vdots \\ t_1 & \cdots & t_p & & & 0 & \cdots & 0 \end{bmatrix}.$$

Here H is a Hermitian circulant and D is a matrix of low rank $r = 2p$. Clearly,

$$D = V \begin{bmatrix} 0 & R \\ R^{\mathrm{H}} & 0 \end{bmatrix} V^{\mathrm{H}} \tag{61}$$

where

$$R = \begin{bmatrix} t_p & t_{p-1} & \cdots & t_1 \\ & t_p & \cdots & t_2 \\ & & \ddots & \vdots \\ & & & t_p \end{bmatrix} \in \mathbb{C}^{p \times p}$$

is upper triangular and

$$V = \begin{bmatrix} I_p & 0 \\ 0 & 0 \\ 0 & I_p \end{bmatrix} \in \mathbb{R}^{n \times r}.$$

Hence we obtain a modified eigenvalue problem of the form (5). The special case where $\tilde{H}$ is tridiagonal with additional entries in the (1,n)- and (n,1)-corners has been considered by Björck and Golub [5].

Since it is very easy to compute the eigenvalues and vectors of circulant matrices [9, p. 72] it may be a successful approach to calculate first the spectral decomposition of H and then treat $\tilde{H}x = \tilde{\lambda}x$ as a modified eigenvalue problem.

The knowledge of the eigenvalues and vectors of banded Toeplitz matrices is of considerable interest in in many applications, especially signal processing.

Note that the eigenvalues of Δ are given by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > -\sigma_p \geq \cdots \geq -\sigma_1$ where the σ_i's are the singular values of R. Thus the interlacing property (28) holds with $r_+ = r_- = p$

Remark. The eigenvalue problem $\tilde{H}x = \tilde{\lambda}x$ can also be treated as the restriction of the eigenvalue problem

$$\hat{H}x = \hat{\lambda}x, \qquad \hat{H} \in \mathbb{C}^{(n+p)\times(n+p)} \tag{62}$$

restricted to

$$Q^{\mathrm{T}}x = [0 \quad I_p]\, x = 0, \quad Q \in \mathbb{R}^{(n+p)\times p}. \tag{63}$$

The matrix $\hat{H}$ has the same form as H in (62) but is of order $n + p$. Restricted eigenvalue problems can be treated very similarly as low rank modified eigenvalue problems [2]. Instead of the matrix $Z(\lambda)$ in (22) one has to analyse

$$\hat{Z}(\lambda) = Q^{\mathrm{T}}(\hat{H} - \lambda)^{-1}Q \tag{64}$$

in the present case. Here it is probably advantageous to consider (60) as a restricted eigenvalue problem since the order of $\hat{Z}$ in (64) is only p.

6. On the modified unitary eigenvalue problem. Let $U \in \mathbb{C}^{n\times n}$ be a matrix with known spectral decomposition

$$U = QTQ^{\mathrm{H}}, \quad T = \mathrm{diag}(\tau, \cdots, \tau_n), \quad Q \text{ unitary}, \tag{65}$$

where the eigenvalues τ_i are arranged so that $0 \leq \arg(\tau_1) \leq \cdots \leq \arg(\tau_n) < 2\pi$.

Let S be a unitary matrix such that $I - S$ has small rank r. We consider in this section the modified unitary eigenvalue problem

$$\tilde{U}x = S^{\mathrm{H}}Ux = \tilde{\tau}x. \tag{66}$$

Equation (66) has to be reformulated for the results of §2 to be applicable. Before doing this we investigate how a unitary S of the above kind must look. To that end we state Lemma 6.1.

LEMMA 6.1. *Let $S \in \mathbb{C}^{n\times n}$ be unitary such that* $\mathrm{rank}(I - S) = r < n$. *Then S can be represented in the form*

$$S = I - X\Theta X^{\mathrm{H}}, \tag{67}$$

where $X \in \mathbb{C}^{n\times r}$ has orthonormal columns and $\Theta = \mathrm{diag}(\theta_1, \cdots, \theta_r)$. The diagonal entries of Θ satisfy the equation

$$|\theta_j - 1|^2 = 1, \quad \theta_j \neq 0. \tag{68}$$

Proof. $I - S$ is normal and thus unitarily diagonalizable. Omitting the trivial eigenvalues and corresponding eigenvectors we obtain the representation

$$I - S = X\Theta X^{\mathrm{H}}, \quad X \in \mathbb{C}^{n\times r}, X^{\mathrm{H}}X = I, \quad \Theta = \mathrm{diag}(\theta_1, \cdots, \theta_r). \tag{69}$$

From $S^{\mathrm{H}}S = I$ we immediately get

$$\Theta + \Theta^{\mathrm{H}} = \Theta^{\mathrm{H}}\Theta, \tag{70}$$

which is an alternate form of (68). □

Remark. Householder matrices are special cases of (67)–(68) with $r = 1$ and $\Theta = 2$.

In order to use the results deduced in §2, we apply the Cayley transform [12, p. 287] on the matrices U and $\tilde{U}$. This transform bijectively maps the set of unitary matrices not having the number 1 in their spectrum on the set of Hermitian matrices. We therefore have to make the assumption that $1 \notin \lambda(U)$ and $1 \notin \lambda(\tilde{U})$. Then the Cayley transform yields the Hermitian matrices

$$H := i(I - U)^{-1}(I + U) \tag{71}$$

and

$$\tilde{H} := i(I - \tilde{U})^{-1}(I + \tilde{U}) = i(S - U)^{-1}(S + U) \tag{72}$$

which have the real eigenvalues $\lambda_1 = i(1+\tau_n)/(1-\tau_n) \geq \cdots \geq \lambda_n = i(1+\tau_1)/(1-\tau_1)$ and $\tilde{\lambda}_1 = i(1 + \tilde{\tau}_n)/(1 - \tilde{\tau}_n) \geq \cdots \geq \tilde{\lambda}_n = i(1 + \tilde{\tau}_1)/(1 - \tilde{\tau}_1)$, respectively. The eigenvectors remain unchanged.

The assumption that neither $\lambda(U)$ nor $\lambda(\tilde{U})$ contains the number 1 is not restrictive since premultiplication of U or S^{H} by a number $a = e^{i\phi}$ turns the spectra on the unit sphere by the angle ϕ. Therefore, choosing ϕ properly, the assumptions can easily be satisfied.

Now we have

$$\begin{aligned}(S - U)^{-1}(S + U) &- (I - U)^{-1}(I + U)\\ &= (I - U)^{-1}[(I - S)(S - U)^{-1}(S + U) - (I - S)]\\ &= 2(I - U)^{-1}(I - S)(S - U)^{-1}U\\ &= -2(I - U)^{-1}(I - S)[I - (I - U)^{-1}(I - S)]^{-1}(I - U^{\mathrm{H}})^{-1}.\end{aligned} \tag{73}$$

The matrix $I-(I-U)^{-1}(I-S)$ with $I-S = X\Theta X^{\mathrm{H}}$ is a straightforward generalization of a Householder elementary matrix [17, p. 3] and thus its inverse (if it exists) is known to be

$$[I - (I - U)^{-1}X\Theta X^{\mathrm{H}}]^{-1} = I - (I - U)^{-1}X\Theta B X^{\mathrm{H}} \tag{74}$$

with

$$B = (I - X^{\mathrm{H}}(I - U)^{-1}X\Theta)^{-1} \in \mathbb{C}^{r\times r}.$$

Setting $V := (I - U)^{-1}X$, one easily obtains

$$\begin{aligned}D = \tilde{H} - H &= \frac{i}{2}[(S - U)^{-1}(S + U) - (S^{\mathrm{H}} + U^{\mathrm{H}})(S^{\mathrm{H}} - U^{\mathrm{H}})^{-1}]\\ &\quad -\frac{i}{2}[(I - U)^{-1}(I + U) - (I + U^{\mathrm{H}})(I - U^{\mathrm{H}})^{-1}]\\ &= iV[\Theta^{\mathrm{H}} - \Theta - \Theta X^{\mathrm{H}}V\Theta B + B^{\mathrm{H}}\Theta^{\mathrm{H}}V^{\mathrm{H}}X\Theta^{\mathrm{H}}]V^{\mathrm{H}}\end{aligned} \tag{75}$$

which is a representation of D in the form (4). Therefore the results of §2 are applicable.

We now derive a divide and conquer algorithm for the eigenvalue problem of a *unitary Hessenberg matrix*, say C. Let $G_1 := I - 2w_1 w_1^{\mathrm{H}}$ be the Householder matrix that maps the first column of C on $\kappa_1 e_1$, $|\kappa_1| = 1$, a multiple of the first unit vector. Then

$$\bar{\kappa}_1 G_1 C = \begin{bmatrix} 1 & 0^{\mathrm{H}} \\ 0 & C_1 \end{bmatrix},$$

where C_1 is a unitary Hessenberg matrix of order $n-1$. Thus, recursively one obtains

$$C = \kappa_1 G_1 \cdots \kappa_n G_n = \kappa G_1 \cdots G_n, \quad \kappa = \prod_{j=1}^{n} \kappa_j, \tag{76}$$

since the G_j are Hermitian and $G_n = I$. Note that the vectors w_j which define the Householder reflectors $I - w_j w_j^{\mathrm{H}}$ have at most *two* nonvanishing components. (For a similar decomposition of C with Givens rotations see [16].)

Without loss of generality we can assume that $\kappa = 1$, which amounts to replacing $\bar{\kappa}C$ by C. Let us furthermore assume that we know the spectral decomposition of $C_L := \prod_{j<k} G_j$ and $C_R := \prod_{j>k} G_j$. C_L and C_R are block diagonal matrices, each with a Hessenberg and an identity block. Since $C = C_L G_k C_R$ is similar to

$$\tilde{U} := SU := G_k C_R C_L, \quad S = G_k, U = C_R C_L \tag{77}$$

we have to compute the eigenvalues of a matrix decomposed in the form (66). The modification S in the present case is such that $I - S$ has rank 1, while U is blockdiagonal with upper Hessenberg blocks of order k and $n-k$, respectively. From (75) we obtain

$$\begin{aligned} D &= 4iv(\bar{\beta} v^{\mathrm{H}} w_k - \beta w_k^{\mathrm{H}} v) v^{\mathrm{H}}, \qquad \beta = 1/(1 - 2w_k^{\mathrm{H}} v) \\ &= 8\,\mathrm{Re}(\beta w_k^{\mathrm{H}} v) v v^{\mathrm{H}} \end{aligned} \tag{78}$$

with

$$v = (I-U)^{-1} w_k = Q(I-T)^{-1} Q^{\mathrm{H}} w_k = \sum_{j=1}^{n} (1-\tau_j)^{-1} (q_j^{\mathrm{H}} w_k) q_j.$$

Note that Q is blockdiagonal, too. The scalar $Z(\lambda)$ in (22) in the present case becomes

$$\begin{aligned} Z(\lambda) &= 8\,\mathrm{Re}(\beta w_k^{\mathrm{H}} v) - v^{\mathrm{H}} Q(\lambda - \Lambda)^{-1} Q^{\mathrm{H}} v \\ &= 8\,\mathrm{Re}(\beta w_k^{\mathrm{H}} v) - v^{\mathrm{H}} Q[\lambda I - i(I-T)^{-1}(I+T)]^{-1} Q^{\mathrm{H}} v \\ &= 8\,\mathrm{Re}(\beta w_k^{\mathrm{H}} v) - w_k^{\mathrm{H}} Q f(\lambda, T) Q^{\mathrm{H}} w_k \end{aligned} \tag{79}$$

with

$$f(\lambda, t) = \frac{1}{\lambda |1-t|^2 + 2\,\mathrm{Im}(t)}. \tag{80}$$

Since T is diagonal, $f(\lambda, T)$ in (79) is evaluated without difficulty.

Remark. An alternative approach to problem (66) can be made if S^{H} and U *commute*. We may then define H by

$$e^{iH} = U, \quad 0 \le H < 2\pi. \tag{81}$$

By (65) we obtain $H = Q\Lambda Q^{\mathrm{H}}$, with $\Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_r)$. The diagonal entries λ_j satisfy the equation $\lambda_j = -i \log \tau_j = \arg \tau_j$. Analogously we define D by

$$e^{iD} = S^{\mathrm{H}}, \quad 0 \leq D < 2\pi. \tag{82}$$

By (67) we then get $D = X \operatorname{diag}(\delta_1, \cdots, \delta_r) X^{\mathrm{H}}$, with $e^{i\delta_j} = 1 - \bar{\theta}_j$, $0 \leq \delta_j < 2\pi$. Now we have

$$S^{\mathrm{H}} U = e^{iD} e^{iH} = e^{i\tilde{H}} \tag{83}$$

with

$$\tilde{H} = H + D. \tag{84}$$

(82) holds if and only if S^{H} and U commute [12]. The advantage of this approach is that both the assumptions $1 \notin \lambda(U)$ and $1 \notin \lambda(\tilde{U})$ are not necessary.

The interlacing property (28) of the eigenvalues of the transformed matrices H and $\tilde{H}$ can be translated in an interlacing property on the unit circle for the eigenvalues of the original matrices U and $\tilde{U}$. By

$$\tau_{j-r_-} \leq \tilde{\tau}_j \leq \tau_{j+r_+} \tag{85}$$

we mean that $\tilde{\tau}_j$ lies on the arc of the unit circle which is passed, when moving from τ_{j-r_-} to τ_{j+r_+} counterclockwise. If $j + r_+ > n$, we identify τ_{j+r_+} with τ_{j+r_+-n}. Likewise we identify τ_{j-r_-} with τ_{j-r_-+n} if $j - r_- < 1$.

Remark. In the case where U is a real orthogonal matrix, it is possible to complete the computations in the real field.

7. Numerical considerations. As mentioned after (25) it is easy to determine any eigenvalue of the modified matrix $\tilde{A}$ at any desired accuracy by bisection: Let $k \in \mathbb{N}, 1 \leq k \leq n, a, b \in \mathbb{R}$ such, that $a \leq \tilde{\lambda} \leq b$ and let $\epsilon > 0$. The following algorithm determines a number $\hat{\lambda}$ satisfying $|\hat{\lambda} - \tilde{\lambda}_k| \leq \epsilon/2$.

ALGORITHM 7.1.

while $b - a > \epsilon$ **do**
begin $c := (a + b)/2$;
  Compute $Z(c)$;
  Determine the inertia $(\pi(Z(c)), \varsigma(Z(c)), \nu(Z(c)))$ of $Z(c)$;
  if $P(c) + \pi(\Delta) - \pi(Z(c)) < k$ **then** $b := c$ **else** $a := c$
end;
$\hat{\lambda} := (a + b)/2$;

To determine starting values for a and b we can skip through the eigenvalues of A and apply (16) to find a $j \in \mathbb{N}$ such that $\lambda_j < \tilde{\lambda}_k < \lambda_{j+1}$, use inequality (28), or both together. Using (28) alone has in general the disadvantage that for each guess c one has to check if c is in $\lambda(A)$.

The iteration in Algorithm 7.1 can eventually be abbreviated if one checks in the case $P(c) + \pi(\Delta) - \pi(Z(c)) \geq k$ if the additional inequality $P(c) + \pi(\Delta) - \pi(Z(c)) - \varsigma(Z(c)) + 1 \leq k$ is satisfied. If the latter holds, we have $c = \tilde{\lambda}_k$.

Since bisection algorithms converge only linearly, it may be desirable to accelerate the iteration. Beattie and Fox [4] derived an inclusion theorem for the eigenvalues of $\tilde{H}$ using the smallest eigenvalue of $Z(\lambda)$. Using further information they have been able to prove an exclusion theorem which is interesting because it gives an interval

centered at λ that does *not* contain an eigenvalue of $\tilde{H}$, thus making it possible to shrink the intervals obtained by bisection even more. It is, however, questionable if the additional labor for the computation of this bound is justified since the rate of convergence remains linear.

Remark. It is worthwhile noting that the above mentioned inclusion theorem has already been used by Rutishauser to locate the smallest eigenvalue of a symmetric matrix. The information obtained made it possible to derive a cubically convergent LR-algorithm [21].

Obviously a higher rate of convergence is obtained if a superlinearly convergent zerofinder is applied to $\det Z(\lambda) = 0$ as soon as a sufficiently small neighbourhood of $\tilde{\lambda}_k$ has been found. But instead of finding a root of $\det Z(\lambda) = 0$ it is more economical to solve for $d_i(\lambda) = 0$, where d_i is a certain element of the diagonal matrix D stemming from the LDL^{T} decomposition [14, p. 84] of $Z(\lambda)$ subjected to a proper permutation. To that end we state the following

THEOREM 7.1. *Let $\lambda \notin \lambda(A)$ and $P \in \mathbb{R}^{r\times r}$ be an arbitrary but fixed permutation matrix. Let*

$$L(\lambda)D(\lambda)L(\lambda)^{\mathrm{H}} = P^{\mathrm{T}}Z(\lambda)P \tag{86}$$

be the so-called LDL^{T} decomposition of $P^{\mathrm{T}}ZP$ where L is a unit lower triangular and D a diagonal matrix. Let $\epsilon > 0$ be such that for all $\lambda \in U_{\lambda^*} := \{\lambda \in \mathbb{R} \,|\, |\lambda - \lambda^*| < \epsilon\}$ the elements of $L(\lambda), D(\lambda)$ and $Z(\lambda)$ are bounded and P is invariant. Then

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} d_i(\lambda) > 0, \qquad 1 \le i \le n, \quad \mu \in U_\lambda^*, \tag{87}$$

where $d_i(\lambda) = e_i^{\mathrm{T}} D(\lambda) e_i$.

Proof. Let $\lambda \in U_{\lambda^*}$. Since

$$Z(\lambda) = \Delta^{-1} - V^{\mathrm{H}}(\lambda - A)^{-1}V = PL(\lambda)D(\lambda)(PL(\lambda))^{\mathrm{H}},$$

we have

$$d_i(\lambda) = e_i^{\mathrm{T}} L(\lambda)^{-1} P^{\mathrm{T}} Z(\lambda) P L(\lambda)^{-\mathrm{H}} e_i.$$

Shortly writing $'$ for $\frac{\mathrm{d}}{\mathrm{d}\lambda}$, we thus obtain

$$\begin{aligned} d_i'(\lambda) \;=\; & e_i^{\mathrm{T}}(L(\lambda)^{-1})' P^{\mathrm{T}} Z(\lambda) P L(\lambda)^{-H} e_i + e_i^{\mathrm{T}} L(\lambda)^{-1} P^{\mathrm{T}} Z'(\lambda) P L(\lambda)^{-\mathrm{H}} e_i \\ & + e_i^{\mathrm{T}} L(\lambda)^{-1} P^{\mathrm{T}} Z(\lambda) P (L(\lambda)^{-\mathrm{H}})' e_i. \end{aligned} \tag{88}$$

Because $(LL^{-1})' = L'L^{-1} + L(L^{-1})' = I' = 0$, the first summand in (88) can be written as

$$-e_i^{\mathrm{T}} L^{-1} L' L^{-1} P^{\mathrm{T}} Z P L^{-\mathrm{H}} e_i = -e_i^{\mathrm{T}} L^{-1} L' D e_i = -(L^{-\mathrm{H}} e_i)^{\mathrm{H}} L^{-1} L' D e_i.$$

Now, $L^{-\mathrm{H}} e_i$ is a vector whose components are nonzero only for components with index $\ge i$ while the first i components of $L'De_i$ vanish since L is *unit* lower triangular. Thus the first and third summand in (88) are zero and therefore

$$\begin{aligned} d_i'(\lambda) &= e_i^{\mathrm{T}} L(\lambda)^{-1} P^{\mathrm{T}} Z'(\lambda) P L(\lambda)^{-\mathrm{H}} e_i \\ &= e_i^{\mathrm{T}} L(\lambda)^{-1} P^{\mathrm{T}} V^{\mathrm{H}} (\lambda - A)^{-2} V P L(\lambda)^{-\mathrm{H}} e_i \\ &= \|(\lambda - A)^{-1} V P L(\lambda)^{-\mathrm{H}} e_i\|_2^2 > 0. \end{aligned} \tag{89}$$

□

Remark. Similarly one obtains

$$d_i''(\lambda) = -2e_i^{\mathrm{T}} L(\lambda)^{-1} P^{\mathrm{T}} V^{\mathrm{H}} (\lambda - A)^{-3} V P L(\lambda)^{-\mathrm{H}} e_i. \tag{90}$$

This theorem is very interesting from the numerical point of view since it states that we can apply the above-mentioned zerofinder on a function with a *simple* root whatever the multiplicity of the eigenvalue sought is! It is not yet clear, however, what happens in the neighbourhood of clustered eigenvalues.

The assumptions made in Theorem 7.1 are satisfied if a stable pivoting strategy is persued. Sehmi [22] stated (89) for the last diagonal element d_r of D for the neighbourhood of a simple eigenvalue of $\tilde{A}$. In the cases studied by Sehmi the matrix $Z(\lambda)$ was diagonally dominant and hence a diagonal pivoting strategy sufficed in the LDL^{T} decomposition of $Z(\lambda)$. This means simply choosing the greatest diagonal element in modulus as pivot. Simpson [23] proved that $d_i'(\lambda) \geq 0$ under less restrictive assumptions.

Thus, if a stable pivoting strategy [6] is combined with diagonal pivoting we can expect that the smallest nontrivial elements in modulus of D will be at the end of its diagonal and it is therefore reasonable to compute the zeros of $d_r(\lambda)$. This has the further advantage that the derivative of $d_r(\lambda)$ is easily computed. By (89) we get

$$d_r'(\lambda) = \|(\lambda - A)^{-1} V P L(\lambda)^{-\mathrm{H}} e_r\|_2^2 = \|(\lambda - A)^{-1} V P e_r\|_2^2, \tag{91}$$

i.e., $d_r'(\lambda)$ is the 2–norm of one of the columns of the matrix $(\lambda - A)^{-1}V$, which has to be formed for the computation of $Z(\lambda)$. Thus we are led to the following algorithm which determines a number $\hat{\lambda}$ that satisfies $|\hat{\lambda} - \lambda_k| \leq \epsilon$ by the Newton iteration method which possesses a locally quadratic order of convergence.

ALGORITHM 7.2.

$c := (a + b)/2;$
while $b - c > \epsilon$ **and** $c - a > \epsilon$ **do**
begin $W := (c - A)^{-1} V$; $Z(c) := V^{\mathrm{H}} W$;
  Compute $L(c)D(c)L(c)^{\mathrm{H}} = P^{\mathrm{T}} Z(c) P$ using diagonal pivoting;
  Determine the inertia of $D(c)$;
  if $P(c) + \pi(\Delta) - \pi(Z(c)) < k$ **then** $b := c$ **else** $a := c$
  $d_r' := \|W P e_r\|_2^2$;
  $c := c - d_r / d_r'$;
  if $c < a$ **or** $c > b$ **then** $c := (b + a)/2$
end;
$\hat{\lambda} := c$;

Acknowledgement. The authors would like to thank Dr. Nav Sehmi and Dr. Dan Sorensen for valuable discussions.

REFERENCES

[1] P. ARBENZ, *On the solution of perturbed and restricted eigenvalue problems of selfadjoint operators by the Weinstein–Aronszajn methods*, unpublished manuscript, December 1986.

[2] P. ARBENZ, W. GANDER AND G. H. GOLUB, *Restricted rank modification of the symmetric eigenvalue problem: theoretical considerations*, Research Report No. 87-01, Seminar für Angewandte Mathematik, ETH Zürich, January 1987.

[3] C. BEATTIE, *An extension of Aronszajn's rule: slicing the spectrum for intermediate problems*, SIAM J. Numer. Anal., 24 (1987), pp. 828–843.

[4] C. BEATTIE AND D. FOX, *Schur complements and the Weinstein–Aronszajn theory for modified matrix eigenvalue problems*, Tech. Report UMSI 87/11, University of Minnesota Supercomputer Institute, February 1987.
[5] Å. BJÖRCK AND G. H. GOLUB, *Eigenproblems for matrices associated with periodic boundary conditions*, SIAM Rev., 19 (1973), pp. 5–16.
[6] J. R. BUNCH AND L. KAUFMAN, *Some stable methods for calculating inertia and solving symmetric linear systems*, Math. Comp., 31 (1977), pp. 163–179.
[7] J. R. BUNCH, C. P. NIELSON AND D. C. SORENSEN, *Rank-one modification of the symmetric eigenproblem*, Numer. Math., 31 (1978), pp. 31–48.
[8] J. J. M. CUPPEN, *A divide and conquer method for the symmetric tridiagonal eigenproblem*, Numer. Math., 36 (1981), pp. 177–195.
[9] P. J. DAVIS, *Circulant Matrices*, Wiley, New York, 1979.
[10] J. J. DONGARRA AND D. C. SORENSEN, *A fully parallel algorithm for the symmetric eigenvalue problem*, SIAM J. Sci. Stat. Comput., 8 (1987), pp. s139–s154.
[11] K. FAN, *Maximum properties and inequalities for the eigenvalues of completely continuous operators*, Proc. Nat. Acad. Sci., 37 (1951), pp. 760–766.
[12] F. R. GANTMACHER, *Matrizentheorie*, 2nd ed., Springer-Verlag, New York, Berlin, 1986.
[13] G. H. GOLUB AND W. KAHAN, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal., Ser. B, 2 (1965), pp. 205–224.
[14] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.
[15] S. H. GOULD, *Variational methods for eigenvalue problems*, 2nd ed., University of Toronto Press, Toronto, 1966.
[16] W. B. GRAGG, *The* QR *algorithm for unitary Hessenberg matrices*, J. Comput. Appl. Math., 16 (1986), pp. 1–8.
[17] A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Dover, New York, 1974.
[18] A. K. JAIN, *Fast inversion of banded Toeplitz matrices by circular decompositions*, IEEE Trans. Acoust. Speech Signal Process., 26 (1978) pp. 121–126.
[19] E. R. JESSUP AND D. C. SORENSEN, *A parallel algorithm for computing the singular value decomposition of a matrix*, unpublished manuscript, March 1987.
[20] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, NJ, 1980.
[21] H. RUTISHAUSER, *Über eine kubisch konvergente variante der lr-transformation*, Z. Angew. Math. Mech., 40 (1960), pp. 49–54.
[22] N. S. SEHMI, *A Newtonian procedure for the solution of the Kron characteristic value problem*, J. Sound Vibration, 100 (1985), pp. 409–421.
[23] A. SIMPSON, *Scanning Kron's determinant*, Quart. J. Mech. Appl. Math., 27 (1974), pp. 27–43.
[24] ———, *The Kron methodology and practical algorithms for eigenvalue, sensitivity and response analyses of large scale structural systems*, Aeronaut. J., 84 (1980), pp. 417–433.
[25] R. C. THOMPSON, *The behavior of eigenvalues and singular values under perturbations of restricted rank*, Linear Algebra Appl., 13 (1976), pp. 69–78.
[26] H. F. WEINBERGER, *Variational Methods for Eigenvalue Approximation*, Regional Conference Series in Applied Mathematics 15, SIAM, Philadelphia, 1974.
[27] A. WEINSTEIN AND W. STENGER, *Methods of Intermediate Problems for Eigenvalues*, Academic Press, New York, 1972.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 1, January 1988

005

SUPERFAST SOLUTION OF REAL POSITIVE DEFINITE TOEPLITZ SYSTEMS*

GREGORY S. AMMAR† AND WILLIAM B. GRAGG‡

Abstract. We describe an implementation of the generalized Schur algorithm for the superfast solution of real positive definite Toeplitz systems of order $n + 1$, where $n = 2^r$. Our implementation uses the split-radix Fast Fourier Transform algorithms for real data of Duhamel. We are able to obtain the nth Szegö polynomial using less than $8n \log_2^2 n$ real arithmetic operations without explicit use of the bit-reversal permutation. Since Levinson's algorithm requires slightly more than $2n^2$ operations to obtain this polynomial, we achieve crossover with Levinson's algorithm at $n = 256$.

Key words. Toeplitz matrix, Schur's algorithm, split-radix Fast Fourier Transform

AMS(MOS) subject classifications. 65F05, 65E05

1. Introduction.

Consider the linear system of equations $Mx = b$, where

$$M = M_{n+1} = \begin{bmatrix} \mu_0 & \mu_1 & \mu_2 & \cdots & \mu_n \\ \mu_1 & \mu_0 & \mu_1 & \cdots & \mu_{n-1} \\ \mu_2 & \mu_1 & \mu_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \mu_1 \\ \mu_n & \mu_{n-1} & \cdots & \mu_1 & \mu_0 \end{bmatrix} = [\mu_{i-j}]_{i,j=0}^{n}$$

is a real symmetric positive definite Toeplitz matrix of order $n + 1$. In contrast with the standard Gaussian and Choleski factorization techniques, which require $O(n^3)$ arithmetic operations, there are several well-known *fast*, $O(n^2)$, methods for solving a Toeplitz system of equations [21], [29], [4], [17]. More recently, several $O(n \log^2 n)$ methods have been presented [6], [8] [12], [22], [20]; we refer to these methods as *superfast* Toeplitz solvers because they require substantially less computation than the fast Toeplitz solvers for sufficiently large n.

It is well known (see, e.g., [19], [18], [3]) that fast Toeplitz solvers are based on ideas from the classical theory of polynomials orthogonal on the unit circle (Szegö polynomials). In particular, the Szegö polynomials can be identified with the columns of the reverse Choleski factorization of M^{-1}. This leads to the observation that the classical Szegö recursions [28], [1], [14] are equivalent with the Levinson–Durbin algorithm for the Yule–Walker equations [16]. Moreover, the decomposition of M^{-1} given by the Gohberg–Semencul formula is equivalent with the Christoffel–Darboux–Szegö formula. Schur's algorithm [23] provides another connection between Toeplitz solvers and classical analysis. Schur's algorithm generates a continued fraction representation of a holomorphic function mapping the unit disk in the complex plane into its closure, and is known to be closely related with the fast algorithms for finding the Choleski factorization of the positive definite Toeplitz matrix M [18], [22].

* Received by the editors August 28, 1987; accepted for publication October 1, 1987. This research was supported in part by the National Science Foundation under grant DMS-8704196. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115.

‡ Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506. Present address, Department of Mathematics, Naval Postgraduate School, Monterey, California 93943. The research of this author was supported in part by the Bergen Scientific Centre, IBM.

A presentation of the superfast algorithm of de Hoog [12] and Musicus [22] that uses the theory of orthogonal polynomials on the unit circle is given in [2], [3]. This algorithm is naturally described in terms of a generalization of Schur's classical algorithm. The generalized Schur algorithm is a doubling procedure for calculating the linear fractional transformation that results from n steps of Schur's algorithm. This formulation provides a concise and classically motivated presentation of the algorithm of de Hoog and Musicus when applied to a positive definite matrix.

The implementation of the generalized Schur algorithm for the superfast solution of a (Hermitian) positive definite Toeplitz system is described in [2]. By using standard Fast Fourier Transform (FFT) techniques to perform the required polynomial recursions, we can construct the linear fractional transformation that results from n steps of Schur's algorithm in $O(n \log_2^2 n)$ complex multiplications. This process yields, without extra work, all n *Schur parameters*, also known as *reflection coefficients* or *partial correlation coefficients*. These parameters are often needed in applications.

The de Hoog–Musicus algorithm consists of two phases. First the nth degree Szegö polynomial is contructed from the linear fractional transformation obtained by the generalized Schur algorithm. Second, the Gohberg–Semencul formula is used to solve the Toeplitz system in $O(n \log_2 n)$ additional multiplications. Each phase involves the computation of cyclic convolutions. These convolutions are performed using in-place FFTs without explicit use of the bit-reversal permutation by using "dual codes" and leaving all transformed data in bit-reversed order. If we insist that the transformed data be in correct order, the number of necessary data accesses increases. Our implementation of the algorithm uses $2\, n \log_2^2 n + O(n \log_2 n)$ complex multiplications [2]. This operation count is less than those obtained by de Hoog and Musicus. Moreover, this algorithm requires the least amount of computation among the other superfast Toeplitz solvers [6], [8], [20].

In this paper we describe an implementation of the generalized Schur algorithm for a real positive definite Toeplitz matrix. The implementation of this superfast Toeplitz solver for real (symmetric) positive definite matrices is conceptually the same. The essential difference is the use of FFT algorithms that exploit the inherent symmetries of the real data and their transforms. There are various ways to perform an FFT on real data in roughly half the computation as in the complex case [27]. We desire the most efficient algorithms possible since transforms of various size need to be performed repeatedly during the algorithm. We also want to be able to perform the real convolutions without explicit use of the bit-reversal permutation. In § 2, we consider some of the real FFT algorithms and show how the real split-radix FFT of Duhamel [13], [24], [25] suits our purpose. The generalized Schur algorithm is described in § 3, and in § 4 its implementation for real input data is described. In § 5 we consider the superfast solution of a real positive definite Toeplitz system of equations by using the generalized Schur algorithm. We will see that the nth degree Szegö polynomial can be calculated in less than $8n \log_2^2 n$ total real operations.

2. Evaluation of real cyclic convolutions. The efficient implementation of the generalized Schur algorithm relies on the use of FFTs to evaluate cyclic convolutions. Several methods exist for calculating the Fourier transform of real data in roughly half the computation of the complex case. Each of these methods yields an efficient method for evaluating real convolutions, and each results in an implementation of the generalized Schur algorithm for real data that requires roughly half the computation as in the complex case. Since convolutions of various sizes are performed repeatedly in the algorithm, we desire the most efficient real transforms possible. Moreover, we want to implement the

algorithm without explicit use of the bit-reversal permutation. We avoided this permutation in the complex case, but since the transform of a real vector is not real, some additional considerations must be made to avoid the bit reversal for the case of real input data.

Recently, real FFT algorithms that can serve as dual codes to allow us to avoid the shuffling have been presented [9], [13], [24], [25]. In this section we show how the split-radix FFT for real data suits our purpose. The algorithms are described by considering the splitting in matrix notation, and precise operation counts are given for use in the subsequent sections. We will need the following lemmas in our derivation. Assume that n and n_0 are powers of two, and let $\lg n := \log_2 n$.

LEMMA 2.1. *If* $\phi(n) = 2\phi(n/2) + 2an \lg n + bn + c$ *for* $n > n_0$, *then*

$$\phi(n) = an \lg^2 n + (a+b)n \lg n + dn - c,$$

where d *is determined by the initial condition* $\phi(n_0)$.

LEMMA 2.2. *If* $\phi(n) = \phi(n/2) + 2\phi(n/4) + an + b$, *then*

$$\phi(n) = \frac{2}{3}an \lg n - \frac{b}{2} + cn + d(-1)^{\nu}$$

where c *and* d *are determined by* $\phi(n_0)$ *and* $\phi(n_0/2)$.

LEMMA 2.3. *If* $\phi(n) = \phi(n/2) + an \lg n + bn + c + d(-1)^{\lg n}$, *then*

$$\phi(n) = 2an \lg n + 2(b-a)n + c \lg n + \frac{d}{2}(-1)^{\lg n} + e$$

where e *is determined by* $\phi(n_0)$.

These lemmas are directly verified by induction and are easily derived by considering the corresponding inhomogeneous linear difference equations for $\phi_\nu := \phi(2^\nu)$.

The *discrete Fourier transform* (*DFT*) of $x \in \mathbb{C}^n$ is defined by $F_n x$, where $nF_n := [\bar{\omega}_n^{jk}]_{j,k=0}^{n-1}$, ω_n is the principal nth root of unity $\exp(2\pi i/n)$, and $\bar{\alpha}$ denotes the complex conjugate of α. The *inverse discrete Fourier transform* (*IDFT*) of $y \in \mathbb{C}^n$ is then given by $W_n y$, where $W_n := F_n^{-1} = n\bar{F}_n = [\omega_n^{jk}]_0^{n-1}$. There are various ways to compute the DFT or IDFT in $O(n \log n)$ arithmetic operations. Such an algorithm is called a *Fast Fourier Transform* (*FFT*). In the following we focus, without loss of generality, on the computation of $y = W_n x$.

Let $K_n = [e_0, e_{n-1}, e_{n-2}, \cdots, e_1]$ and $J_n = [e_{n-1}, e_{n-2}, \cdots, e_0]$, respectively, be the $n \times n$ *reflection* and *reversal matrices*, where $e_0, \cdots, e_{n-1}$ are the columns of the $n \times n$ identity matrix. Then we have

$$K_n W_n = [\omega_n^{-jk}] = \bar{W}_n \tag{2.1}$$

and

$$J_n W_n = [\omega_n^{(n-j-1)k}] = [\omega_n^{-jk}\omega_n^{-k}] = \bar{W}_n \bar{D}_n, \tag{2.2}$$

where $D_n := \operatorname{diag}[\omega_n^j]_0^{n-1}$. It is easily seen from (2.1) that whenever $x \in \mathbb{R}^n$, $y = F_n x$ satisfies $K_n y = \bar{y}$; that is, $\eta_{n-j} = \bar{\eta}_j$ ($j = 1, \cdots, n/2 - 1$) and $\eta_0, \eta_{n/2} \in \mathbb{R}$. We will say the transformed vector y possesses *conjugate-even* (*CE*) *symmetry*. Thus the transform of a real vector is determined by the n real numbers that constitute its first $n/2 + 1$ components. There are various methods to compute the real to CE transform and its inverse in roughly half the computation as in the general (complex) case. Some of these methods are considered below.

Let $\alpha_{\mathbb{C}}$, $\mu_{\mathbb{C}}$, respectively, denote a complex addition and multiplication, and similarly for $\alpha_{\mathbb{R}}$, $\mu_{\mathbb{R}}$. Also let $\tau_{\mathbb{R}}$ denote a real arithmetic operation. We determine the num-

ber of real arithmetic operations using $\alpha_{\mathbb{C}} = 2\alpha_{\mathbb{R}}$, $\mu_{\mathbb{C}} = 2\alpha_{\mathbb{R}} + 4\mu_{\mathbb{R}} = 6\tau_{\mathbb{R}}$. In the following we ignore multiplication by 1 and $i = \sqrt{-1}$. We also count the computation of the product of a complex number with an eighth root of unity as $2\alpha_{\mathbb{R}} + 2\mu_{\mathbb{R}}$ (since, e.g., $\omega_8(\alpha + i\beta) = (\alpha - \beta)/\sqrt{2} + i(\alpha + \beta)/\sqrt{2}$).

Our interest in FFTs is motivated by the desire to efficiently compute products of polynomials, or equivalently, cyclic convolutions. The *cyclic* (or *periodic*) *convolution* $x * y =: z = [\zeta_j]_0^{n-1}$ of $x = [\xi_j]_0^{n-1}$ and $y = [\eta_j]_0^{n-1}$ is defined by $\zeta_j = \sum_{k=0}^{n-1} \xi_k \eta_{j-k}$ (where $\eta_{-k} \equiv \eta_{n-k}$). Note that from this definition, the computation of z requires $O(n^2)$ operations. It is easily verified, however, that $W_n z = (W_n x) \cdot (W_n y)$ (and $F_n z = (F_n x) \cdot (F_n y)$), where $u \cdot v$ denotes the Schur product (componentwise product) of the vectors u and v. Thus, if $\phi(n)$ denotes the computation required to compute a complex FFT of order n, $z = x * y$ can be computed using $3\,\phi(n) + n\mu_{\mathbb{C}}$ operations. Moreover, if $\tau(n)$ denotes the computation required by a real to CE transform or its inverse transform, then the cyclic convolution of x, $y \in \mathbb{R}^n$ can be computed in $3\tau(n) + (n/2 - 1)\mu_{\mathbb{C}} + 2\mu_{\mathbb{R}} (n > 1)$.

The calculation of the Fourier and inverse Fourier transforms is typically performed using the Cooley–Tukey algorithm, which can be described as follows. We describe the inverse transform $y = W_n x$ of $x \in \mathbb{C}^n$; since $nF_n = \bar{W}_n$, the computation of $F_n x$ is completely analogous and involves the same amount of computation. We neglect division by n, which is assumed to be a power of two.

Let $m := n/2$, $D'_m := \operatorname{diag}\,[\omega_n^j]_0^{m-1}$, and define the permutation matrix P_n by $P_n^T x = \begin{bmatrix} x'_0 \\ x'_1 \end{bmatrix}$, where $x'_0 = [\xi_{2j}]_0^{m-1}$ and $x'_1 = [\xi_{2j+1}]_0^{m-1}$ are, respectively, the *even* and *odd parts* of x. Then it is easily seen that

$$\text{(2.3a)} \qquad \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} := y = W_n P_n P_n^T x = \begin{bmatrix} W_m & D'_m W_m \\ W_m & -D'_m W_m \end{bmatrix} \begin{bmatrix} x'_0 \\ x'_1 \end{bmatrix} = \begin{bmatrix} u_0 + u_1 \\ u_0 - u_1 \end{bmatrix},$$

where

$$\text{(2.3b)} \qquad \begin{bmatrix} u_0 \\ u_1 \end{bmatrix} := \begin{bmatrix} W_m x'_0 \\ D'_m W_m x'_1 \end{bmatrix}.$$

Thus a DFT or IDFT can be computed from the transforms of the even and odd parts of the input vector with some local work. The repeated application of this splitting leads to the *Cooley–Tukey* FFT algorithm [26]. This method for the computation of $F_n x$ is often called a *radix-two Cooley–Tukey decimation-in-time algorithm* [7]. (The output is given by two transforms of subsets of the input, and in applications the input typically corresponds with data in the time domain.) We will refer to this procedure for the computation of $F_n x$ or $W_n x$ as a *decimation-in-input* (*DII*) algorithm. The DII algorithm (2.3) can be performed in place (i.e., without the use of a temporary work vector) by first permuting the input vector according to the permutation

$$\Pi_n = P_n \operatorname{diag}\,(P_{n/2}, P_{n/2}) \cdots \operatorname{diag}\,(P_4, P_4, \cdots, P_4).$$

The transformation $\Pi_n x$ is referred to as the *bit-reversal permutation of order n*. Thus, in-place computation of the DII FFT is achieved by rearranging the input vector before the computations take place.

Analogously, the even and odd halves of the output are given by two transforms of order $m = n/2$. This algorithm, which is called the *Sande–Tukey* or the *radix-two Cooley–Tukey decimation-in-frequency* algorithm, is given by the recursive application of the formula

$$\text{(2.4a)} \qquad \begin{bmatrix} y'_0 \\ y'_1 \end{bmatrix} := P_n^T W_n x = \begin{bmatrix} W_m & W_m \\ W_m D'_m & -W_m D'_m \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} W_m u_0 \\ W_m D'_m u_1 \end{bmatrix},$$

where

$$\begin{bmatrix} u_0 \\ u_1 \end{bmatrix} := \begin{bmatrix} x_0 + x_1 \\ x_0 - x_1 \end{bmatrix}. \tag{2.4b}$$

The resulting *decimation-in-output* (*DIO*) method, when implemented in place, receives input in correct order and generates output in bit-reversed order.

Note that both methods require the same amount of computation $\phi(n)$. In particular, we have $\phi(1) = 0$, $\phi(2) = 2\alpha_{\mathbb{C}}$ and

$$\phi(n) = 2\phi(n/2) + n\alpha_{\mathbb{C}} + (n/2 - 2)\mu_{\mathbb{C}}$$

for $n = 2^\nu > 1$. Since two of the entries of D'_m are eighth roots of unity when $n > 4$, we have

$$\phi(4) = 16\alpha_{\mathbb{R}},$$

$$\phi(n) = 2\phi(n/2) + (3n - 4)\alpha_{\mathbb{R}} + (2n - 12)\mu_{\mathbb{R}} \qquad (n > 4).$$

Thus, by Lemma 2.1, the computation of the DFT and IDFT of $x \in \mathbb{C}^n$ by either the DII or DIO radix-two FFT algorithms requires at most

$$\phi(n) = (3n \lg n - 3n + 4)\alpha_{\mathbb{R}} + (2n \lg n - 7n + 12)\mu_{\mathbb{R}} \qquad (n = 2^\nu > 4).$$

In many applications we need the transformed data in correct order, so the explicit use of the bit-reversal permutation is required. However, since we are using the transforms as a computational tool for cyclic convolution evaluation, we do not need to know the true order of the components of the transformed vectors; we need only multiply corresponding components of the transformed vectors and apply the inverse transform. Thus, we can use the above two FFT algorithms as *dual codes*, one for the transform and the other for the inverse transform, in order to avoid the need to shuffle the data before or after the calculations. To calculate $z = x * y$ we use the Sande–Tukey (DIO) recursions to obtain $W_n x$ and $W_n y$ in bit-reversed order, form their Schur product to obtain $W_n z$ in bit-reversed form, and then use the Cooley–Tukey (DII) recursions to obtain z in correct order. (Note that we are using an IDFT as our transform and a DFT as our inverse transform.) While the use of dual FFT codes does not affect the amount of computation in the evaluation of cyclic convolutions, it reduces the number of data accesses.

Now consider the computation of $y = W_n x$ when $x \in \mathbb{R}^n$. Note that the DII recursion (2.3) splits the transform of x into two real transforms of half the size. These transforms also possess CE symmetry, so the redundant computations can be identified (and ignored) during each stage of the splitting. In particular, only $y_0 = [\eta_j]_0^{m-1}$ and η_m are computed from components 0 through $l = n/4$ of u_0 and u_1. The successive application of this splitting for the transform of a real vector is known as the *Edson–Bergland algorithm* [5]. This algorithm is described in a recent paper by Swarztrauber [27], where analogous algorithms for vectors with other types of symmetries are derived.

Letting $\tau(n)$ denote the amount of computation used to compute the DFT or IDFT of a real vector using the Edson–Bergland algorithm, we see that $\tau(1) = 0$, $\tau(2) = 2\alpha_{\mathbb{R}}$, $\tau(4) = 6\alpha_{\mathbb{R}}$, and for $n > 4$

$$\begin{aligned} \tau(n) &= 2\tau(n/2) + 2\alpha_{\mathbb{R}} + (n/4 - 1)(2\alpha_{\mathbb{C}} + \mu_{\mathbb{C}}) \\ &= 2\tau(n/2) + (3n/2 - 4)\alpha_{\mathbb{R}} + (n - 6)\mu_{\mathbb{R}} \end{aligned}$$

since one of the complex multiplications is by an eighth root of unity. The Edson–Bergland algorithm therefore requires at most

$$\tau(n)=\left(\frac{3n\lg n}{2}-\frac{5n}{2}+4\right)\alpha_{\mathbb{R}}+\left(n\lg n-\frac{7n}{2}+6\right)\mu_{\mathbb{R}}$$

operations when $n = 2^\nu > 2$ and $\tau(2) = 2\alpha_{\mathbb{R}}$.

Note that the Edson–Bergland algorithm requires the (real) input data to be in bit-reversed form. In the pursuit of our desire to eliminate the explicit use of the bit-reversal permutation, we want the real data in correct order, and the transforms, having CE symmetry, in bit-reversed order. Such an algorithm can be derived from the DIO recursions applied to a real vector.

Let us consider the DIO splitting (2.4) applied to $x \in \mathbb{R}^n$. In this case $y'_0 = W_m u_0$ is the transform of a real vector, while $y'_1 = W_m D'_m u_1$. A further splitting reveals the redundancies in y'_1. Let $l := n/4 \geqq 1$, $u_1 =: \left[\begin{smallmatrix} t_0 \\ t_1 \end{smallmatrix}\right]$ and $D''_l := \operatorname{diag}\left[\omega_{4l}^j\right]_{j=0}^{l-1}$. Then $(D''_l)^4 = (D'_l)^2 = D_l$, $(D''_l)^{-1} = \bar{D}''_l$, $D'_m = \left[\begin{smallmatrix} D'_l & 0 \\ 0 & iD'_l \end{smallmatrix}\right]$ and

$$\begin{bmatrix} v'_0 \\ v'_1 \end{bmatrix} := P_m^T y'_1 = \begin{bmatrix} W_l D''_l z_0 \\ W_l D'_l D''_l \bar{z}_0 \end{bmatrix} \tag{2.5}$$

where $z_0 = t_0 + it_1$. Furthermore, by (2.2) we have

$$J_l v'_1 = J_l W_l D'_l D''_l \bar{z}_0 = \bar{W}_l \bar{D}''_l \bar{z}_0 = \bar{v}'_0,$$

so the computation of v'_1 is redundant. Thus, y can be calculated using one real FFT of order $n/2$, one complex FFT of order $n/4$ and some local work. This splitting strategy results in the *real split-radix FFT algorithm* of Duhamel [13], [24], [25]. Since the product $D''_l z_0$ requires $l - 1$ complex multiplications, with one eighth root of unity if $l > 1$, the total work $\tau(n)$ for a split-radix FFT on a real vector of order n satisfies

$$\tau(n) = \tau(n/2) + \phi(n/4) + (3n/2 - 2)\alpha_{\mathbb{R}} + (n-6)\mu_{\mathbb{R}} \qquad (n > 4), \tag{2.6}$$

with $\tau(4) = 6\alpha_{\mathbb{R}}$, $\tau(2) = 2\alpha_{\mathbb{R}}$. Note that the corresponding DII complex to real transform (the inverse transform) is also given by the above splitting and requires the same amount of computation.

The split-radix technique can also be applied to the computation of a complex FFT. Let $x \in \mathbb{C}^n$, $y := W_n x$, and split y'_1 in (2.4) to obtain

$$P_m^T y'_1 =: \begin{bmatrix} v'_0 \\ v'_1 \end{bmatrix} = \begin{bmatrix} W_l D''_l z_0 \\ W_l D'_l D''_l \bar{z}_1 \end{bmatrix} \quad \text{where} \quad \begin{bmatrix} z_0 \\ z_1 \end{bmatrix} := \begin{bmatrix} t_0 + it_1 \\ \bar{t}_0 + i\bar{t}_1 \end{bmatrix}.$$

In this case v'_1 is not redundant, but we have $v'_1 = J_l \overline{W_l D''_l z_1}$.

Thus, the computational work $\phi(n)$ for a complex split-radix FFT of order n satisfies

$$\phi(n) = \phi(n/2) + 2\phi(n/4) + (4n-4)\alpha_{\mathbb{R}} + (2n-12)\mu_{\mathbb{R}} \qquad (n > 4),$$

$$\phi(4) = 16\alpha_{\mathbb{R}}, \qquad \phi(2) = 4\alpha_{\mathbb{R}}.$$

We therefore have by Lemma 2.2 the following.

PROPOSITION 2.1. *The computation of a complex* FFT *of size $n = 2^\nu$ by the split-radix method requires at most*

$$\phi(n) = \left(\frac{8}{3}n\nu - \frac{16}{9}n - \frac{2}{9}(-1)^\nu + 2\right)\alpha_{\mathbb{R}} + \left(\frac{4}{3}n\nu - \frac{38}{9}n + \frac{2}{9}(-1)^\nu + 6\right)\mu_{\mathbb{R}}$$

$$= 4n\nu - 6n + 8 \quad \textit{total real operations}$$

for $n \geqq 2$.

Thus, the split-radix procedure obtains the DFT with roughly 80% of the computation of the radix-two method. A proportionate amount of computational savings for the evaluation of a real FFT and its (complex to real) inverse is obtained when the real split-radix FFT is performed using a split-radix complex FFT. In particular, by (2.6), Proposition 2.1, and Lemma 2.3, we obtain the following.

PROPOSITION 2.2. *The computation* $\tau(n)$ *required to compute the* FFT *or* IDFT *of a real vector of order* $n = 2^\nu$ *and the* FFT *or* IDFT *of a vector with* CE *symmetry is at most*

$$\tau(n) = \left(\frac{4}{3}n\nu - \frac{17}{9}n - \frac{1}{9}(-1)^\nu + 3\right)\alpha_{\mathbb{R}} + \left(\frac{2}{3}n\nu - \frac{19}{9}n + \frac{1}{9}(-1)^\nu + 3\right)\mu_{\mathbb{R}}$$
$$= 2n\nu - 4n + 6 \quad \textit{total real operations}$$

for $n \geqq 2$.

These operation counts agree with those reported in [24], [25].

Thus, a real cyclic convolution of order n can be computed without explicit use of the bit-reversal permutation by using the real to complex DIO split-radix FFT and its DII inverse as dual codes. These algorithms can be performed in place using real arithmetic and n real storage locations. Moreover, the split-radix FFT has the smallest operation count among the known FFT algorithms for real transforms of length equal to a power of two [25].

We remark that we investigated the use of other methods for real transforms before becoming aware of the split-radix method. We first considered the use of the common method of evaluating the transform of a real vector by the transform of a complex vector of half the size and postprocessing (see, e.g., [26], [27]). However, this method requires the explicit use of the bit-reversal permutation, and moreover, is not as efficient as the Edson–Bergland algorithm. As we remarked above, the Edson–Bergland algorithm is not appropriate for our use either. We considered the use of Hartley transforms [10], and we in fact initially implemented the algorithm using dual codes for the Hartley transform. However, while it requires less computation than the common pre- and post-processing procedure, the radix-two computation of a Hartley transform requires more computation than the Edson–Bergland algorithm. Moreover, the Hartley transform of a convolution is not equal to the Schur product of Hartley transforms, so slightly more computation is needed to form the transform of the convolution from the transforms of the input data. Direct observation shows that the Fourier transform of $x \in \mathbb{R}^4$ requires $6\alpha_{\mathbb{R}}$ while the Hartley transform requires $8\alpha_{\mathbb{R}}$. We therefore do not expect the use of the Hartley transform to be as efficient as Fourier transforms in the evaluation of real convolutions. We have not considered, however, the relative performance of Fourier and Hartley transforms in the evaluation of real convolutions where the input have additional symmetry (e.g., real even or real odd input vectors).

In our study of the development of DIO real to complex FFT analogous with the Edson–Bergland FFT, we found the procedure described by (2.5), and afterward were made aware that this procedure is in fact the real split-radix FFT of Duhamel, which when applied to the real and complex FFT together, uses 20% less computation than the standard radix-two methods. In the meantime we have also become aware of the paper by Briggs [9], where DIO methods are derived that are analogous with the symmetric DII FFTs presented in [27], including the Edson–Bergland algorithm.

3. The generalized Schur algorithm. Let ϕ be a *Schur function*, which is to say that ϕ is a holomorphic function that maps the open unit disk D in the complex plane into its closure. Schur's algorithm [23] is a procedure that generates a sequence of Schur

functions by the successive application of linear fractional transformations (LFTs) to ϕ. It is defined as follows.

SCHUR'S ALGORITHM.
input: an initial Schur function ϕ_0,
$\gamma_1 := \phi_0(0)$,
for $n = 1, 2, 3, \cdots$ **while** $|\gamma_n| < 1$

$$\phi_n(\lambda) := \frac{1}{\lambda} \frac{\phi_{n-1}(\lambda) - \gamma_n}{1 - \bar{\gamma}_n \phi_{n-1}(\lambda)},$$

$$\gamma_{n+1} := \phi_n(0).$$

If $|\gamma_n| = 1$, then $\phi_{n-1}(\lambda) \equiv \gamma_n$ and Schur's algorithm terminates. On the other hand, if $|\gamma_n| < 1$ then ϕ_n is a Schur function. Thus, Schur's algorithm generates a possibly finite sequence of Schur functions ϕ_n that satisfy $\phi_{n-1} = t_n(\phi_n)$, where

$$t_n(\tau) = (\gamma_n + \lambda\tau)/(1 + \bar{\gamma}_n \lambda\tau).$$

We can therefore view the algorithm as the generation of a continued fraction representation of ϕ_0 (since continued fractions are related with compositions of LFTs). In particular, $\phi_0 = T_n(\phi_n)$, where $T_n = t_1 \circ t_2 \circ \cdots \circ t_n$. The function $T_n(0)$ is referred to as the *nth approximant* of ϕ_0, and ϕ_n is called the *nth tail* of ϕ_0.

In the standard implementation of Schur's algorithm, each ϕ_n is written as a quotient of formal power series,

$$\phi_n =: \frac{\alpha_n(\lambda)}{\beta_n(\lambda)} = \frac{\sum_k \alpha_{n,k} \lambda^k}{\sum_k \beta_{n,k} \lambda^k},$$

with $\beta_n(0) > 0$ as a *partial normalization.* The computations are then arranged so that the coefficient pairs $(\alpha_{0,k}, \beta_{0,k})$ are entered and processed in a sequential manner. This results in the following.

PROGRESSIVE SCHUR ALGORITHM.
for $k = 1, 2, 3, \cdots$ **while** $|\gamma_k| < 1$
enter $\alpha_{0,k-1}, \beta_{0,k-1}$
for $j = 1, 2, 3, \cdots, k-1$

$$\begin{bmatrix} \alpha_{j,k-j-1} \\ \beta_{j,k-j} \end{bmatrix} = \begin{bmatrix} 1 & -\gamma_j \\ -\bar{\gamma}_j & 1 \end{bmatrix} \begin{bmatrix} \alpha_{j-1,k-j} \\ \beta_{j-1,k-j} \end{bmatrix}$$

$\gamma_k = \alpha_{k-1,0}/\beta_{k-1,0}$,
$\beta_{k,0} = \beta_{k-1,0}(1 - |\gamma_k|^2)$.

Of course, in practice a finite number of coefficients are input. Let $\alpha^{(n)}$ denote the polynomial of degree less than n formed by the first n terms of the power series α. If $\alpha_0^{(n)}, \beta_0^{(n)}$ are input, then the progressive Schur algorithm calculates $\alpha_k^{(n-k)}, \beta_k^{(n-k)}$ and γ_k for $k = 1, \cdots, n$, using at most $n^2\alpha_{\mathbb{R}} + n(n+2)\mu_{\mathbb{R}}$.

Schur's algorithm can also be formulated in terms of the LFTs T_n. In particular, it is easily seen that

$$T_n(\tau) = \frac{\xi_n + \tilde{\eta}_n \tau}{\eta_n + \tilde{\xi}_n \tau},$$

where ξ_n, η_n are the polynomials that satisfy the recurrence relations

$$\begin{bmatrix} \xi_n \\ \eta_n \end{bmatrix} = \begin{bmatrix} \tilde{\eta}_{n-1} & \xi_{n-1} \\ \tilde{\xi}_{n-1} & \eta_{n-1} \end{bmatrix} \begin{bmatrix} \gamma_n \\ 1 \end{bmatrix}, \qquad \begin{bmatrix} \xi_0 \\ \eta_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{3.1}$$

where $\tilde{\xi}_n(\lambda) := \lambda^n \bar{\xi}_n(1/\lambda)$ and $\tilde{\eta}_n(\lambda) := \lambda^n \bar{\eta}_n(1/\lambda)$. We refer to ξ_n and η_n as the nth *Schur polynomials* associated with the Schur function ϕ_0. Note that (3.1) implies (for $n \geqq 1$)

$$\deg \xi_n < n, \qquad \deg \eta_n < n,$$

$$\xi_n(0) = \gamma_1, \qquad \eta_n(0) = 1,$$

as well as the *determinant formula*

$$\eta_n \tilde{\eta}_n - \xi_n \tilde{\xi}_n = \delta_n \lambda^n,$$

where $\delta_n = (1 - |\gamma_1|^2) \cdots (1 - |\gamma_n|^2)$.

The generalized Schur algorithm is based on a doubling procedure for generating T_{2n} from T_n. The idea can be described in general terms as follows. Let $\phi_n = T_n^{-1}(\phi_0)$ be the nth tail of ϕ_0, and let $T_{n,n}$ denote the LFT that results from n steps of Schur's algorithm applied to the Schur function ϕ_n. The LFT T_{2n} is then given by the composition $T_n \circ T_{n,n}$.

For each n, let α_n and β_n be formal power series such that $\phi_n = \alpha_n/\beta_n$. The nth tail of ϕ_0 is then given by

$$\phi_n = \frac{\alpha_n}{\beta_n} = T_n^{-1}(\phi_0) = \frac{\alpha_0 \eta_n - \beta_0 \xi_n}{\beta_0 \tilde{\eta}_n - \alpha_0 \tilde{\xi}_n}.$$

It is shown in [2], [3] that both the numerator and denominator in the last expression are formal power series that are divisible by λ^n. In particular,

$$\alpha_0 \eta_n - \beta_0 \xi_n = \gamma_{n+1} \delta_n \lambda^n + O(\lambda^{n+1}),$$

$$\beta_0 \tilde{\eta}_n - \alpha_0 \tilde{\xi}_n = \delta_n \lambda^n + O(\lambda^{n+1}).$$

We can therefore take

$$\alpha_n = (\alpha_0 \eta_n - \beta_0 \xi_n)/\lambda^n, \qquad \beta_n = (\beta_0 \tilde{\eta}_n - \alpha_0 \tilde{\xi}_n)/\lambda^n. \tag{3.2}$$

These formulas constitute the first component of the generalized Schur algorithm: the computation of α_n and β_n (i.e., ϕ_n) from ξ_n, η_n, α_0, and β_0.

The next step in the generalized Schur algorithm is the doubling step. Since $\phi_n = \alpha_n/\beta_n$ is a Schur function we can obtain the nth Schur polynomials $\xi_{n,n}$ and $\eta_{n,n}$ (i.e., $T_{n,n}$) from α_n and β_n using the same procedure that $\xi_{0,n} = \xi_n$ and $\eta_{0,n} = \eta_n$ were obtained from α_0 and β_0.

The third step of the algorithm is the computation of $\xi_{0,2n}$ and $\eta_{0,2n}$ from the composition of T_n and $T_{n,n}$. In particular, for any $k > 0$ we have $\phi_n = T_{n,k}(\phi_{n+k})$; that is, the kth tail of ϕ_n is the $(n + k)$th tail of ϕ_0. We therefore have $\phi_0 = T_{n+k}(\tau) = T_n(T_{n,k}(\tau))$, or equivalently,

$$\xi_{0,n+k} = \tilde{\eta}_{0,n} \xi_{n,k} + \xi_{0,n} \eta_{n,k}, \qquad \eta_{0,n+k} = \tilde{\xi}_{0,n} \xi_{n,k} + \eta_{0,n} \eta_{n,k}. \tag{3.3}$$

Equations (3.2) and (3.3) form the basis of the generalized Schur algorithm. The algorithm is easiest to describe in its recursive form.

GENERALIZED SCHUR ALGORITHM.

input: $n = 2^\nu$ and polynomials $\alpha_0^{(n)}$, $\beta_0^{(n)}$, where α_0, β_0 are power series such that $\phi_0 := \alpha_0/\beta_0$ is a Schur function;

$\xi_{0,1} = \gamma_1 = \alpha_0^{(1)}$, $\eta_{0,1} = 1$;
for $m = 1, 2, 4, \cdots, n/2$
1. use (3.2) to obtain $\alpha_m^{(m)}$, $\beta_m^{(m)}$ from $\xi_{0,m}$, $\eta_{0,m}$, $\alpha_0^{(2m)}$, $\beta_0^{(2m)}$;
2. use the generalized Schur algorithm to compute $\xi_{m,m}$, $\eta_{m,m}$ from $\alpha_m^{(m)}$ and $\beta_m^{(m)}$ as $\xi_{0,m}$ and $\eta_{0,m}$ were obtained from $\alpha_0^{(m)}$ and $\beta_0^{(m)}$;
3. use (3.3) to compute $\xi_{0,2m}$, $\eta_{0,2m}$ from $\xi_{0,m}$, $\eta_{0,m}$, $\xi_{m,m}$, $\eta_{m,m}$;

output: the Schur polynomials $\xi_n = \xi_{0,n}$, $\eta_n = \eta_{0,n}$ and the Schur parameters $[\gamma_j]_1^n$.

Note that the classical Schur algorithm generates $\alpha_k^{(n-k)}$, $\beta_k^{(n-k)}$ using the n Schur parameters γ_k as intermediate values. In the generalized Schur algorithm, however, the number of coefficients of α_k, β_k to be calculated is equal to the largest power of two that divides k. For example, $n/2$ coefficients are calculated when $k = n/2$, $n/4$ coefficients are calculated when $k = n/4$ and $k = 3n/4$, and only the constant terms are calculated when k is odd. Nevertheless, *every Schur parameter is generated in the generalized Schur algorithm.* These parameters are often of physical and mathematical significance; if they are not needed for the output, however, they do not need to be stored in the above algorithm.

4. Implementation of the generalized Schur algorithm for real data. In order to implement the generalized Schur algorithm, we write steps 1 and 3 as convolutions and apply FFT techniques. Let $\mathbb{R}_n[\lambda]$ denote the set of real polynomials of degree less than n. For any polynomial $\xi(\lambda) = \sum_0^{n-1} \xi_j \lambda^j \in \mathbb{R}_n[\lambda]$ we write $x \leftrightarrow \xi$ for the associated vector $x = [\xi_j]_0^{n-1} \in \mathbb{R}^n$. Define vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$ ($m = n/2 \geqq 1$) by

$$x_0, x_1, x \leftrightarrow \xi_m, \xi_{m,m}, \xi_n, \qquad a_0, a_1, a \leftrightarrow \alpha^{(m)}, \alpha_m^{(m)}, \alpha^{(n)}$$

and similarly for $y \leftrightarrow \eta$, $b \leftrightarrow \beta$. Also define $\tilde{x} = J_n \bar{x}$, so that $\tilde{x} \leftrightarrow \tilde{\xi}_n/\lambda$.

Step 3 involves the products of polynomials in $\mathbb{R}_m[\lambda]$ and $\mathbb{R}_{m+1}[\lambda]$, which are equivalent with convolutions of size n. Specifically,

$$x = E_n \begin{bmatrix} \tilde{y}_0 \\ 0 \end{bmatrix} * \begin{bmatrix} x_1 \\ 0 \end{bmatrix} + \begin{bmatrix} x_0 \\ 0 \end{bmatrix} * \begin{bmatrix} y_1 \\ 0 \end{bmatrix},$$

$$y = E_n \begin{bmatrix} \tilde{x}_0 \\ 0 \end{bmatrix} * \begin{bmatrix} x_1 \\ 0 \end{bmatrix} + \begin{bmatrix} y_0 \\ 0 \end{bmatrix} * \begin{bmatrix} y_1 \\ 0 \end{bmatrix},$$

where $E_n = [e_1, e_2, \cdots, e_{n-1}, e_0]$ is the $n \times n$ *cyclic downshift matrix.*

While Step 1 involves polynomials in $\mathbb{R}_{3m}[\lambda]$, only the *middle* m coefficients of these polynomials are needed. We can therefore obtain a_1, b_1 from the last m components of convolutions of order n. In particular,

$$\begin{bmatrix} * \\ a_1 \end{bmatrix} = a * \begin{bmatrix} y_0 \\ 0 \end{bmatrix} - b * \begin{bmatrix} x_0 \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} * \\ b_1 \end{bmatrix} = b * E_n \begin{bmatrix} \tilde{y}_0 \\ 0 \end{bmatrix} - a * E_n \begin{bmatrix} \tilde{x}_0 \\ 0 \end{bmatrix}.$$

We can therefore perform one step of the generalized Schur algorithm, the calculation of $[u, v] := W_n[x, y]$ from a, b and $[u_0, v_0] := W_m[x_0, y_0]$ as follows. The number of real arithmetic operations used follows each substep in brackets.

0. Suppose $[u_0, v_0]$ has been computed $[\omega(m)]$.
1. Compute $[a_1, b_1]$:

   (a) $[x_0, y_0] = F_m[u_0, v_0]$ $[2\tau(m)]$,

   (b) $[p, q] = W_n \begin{bmatrix} x_0 & y_0 \\ 0 & 0 \end{bmatrix}$ $[2\phi(m/2) + (2m - 4)\alpha_{\mathbb{R}} + (4m - 12)\mu_{\mathbb{R}}]$,

   (c) $[r, s] = W_n E_n \begin{bmatrix} \tilde{x}_0 & \tilde{y}_0 \\ 0 & 0 \end{bmatrix}$ $[free]$,

   (d) $[c, d] = W_n[a, b]$ $[2\tau(2m)]$,

   (e) $\begin{bmatrix} * \\ a_1 \end{bmatrix} = F_n(c \cdot q - d \cdot p), \begin{bmatrix} * \\ b_1 \end{bmatrix} = F_n(d \cdot s - c \cdot r)$,
   $[2\tau(2m) + (12m - 8)\alpha_{\mathbb{R}} + (16m - 8)\mu_{\mathbb{R}}]$.

2. Compute $[u_1, v_1]$ from $[a_1, b_1]$ as $[u_0, v_0]$ was computed from $[a_0, b_0]$ $[\omega(m)]$.
3. Computation of $[u, v]$:

   (a) $[x_1, y_1] = F_m[u_1, v_1]$ $[2\tau(m)]$,

   (b) $[p_1, q_1] = W_n \begin{bmatrix} x_1 & y_1 \\ 0 & 0 \end{bmatrix}$ $[2\phi(m/2) + (2m - 4)\alpha_{\mathbb{R}} + (4m - 12)\mu_{\mathbb{R}}]$,

   (c) $u = s \cdot p_1 + p \cdot q_1$, $\quad v = r \cdot p_1 + q \cdot q_1$ $[(12m - 8)\alpha_{\mathbb{R}} + (16m - 8)\mu_{\mathbb{R}}]$.

We compute p, q from u_0, v_0 as follows. We have

$$\begin{bmatrix} p'_0 \\ p'_1 \end{bmatrix} := P_n^T W_n \begin{bmatrix} x_0 \\ 0 \end{bmatrix} = \begin{bmatrix} W_m x_0 \\ W_m D'_m x_0 \end{bmatrix},$$

so that $p'_0 = u_0$ and $p'_1 = W_m D'_m F_m u_0$. Moreover, if $l := n/4 \geqq 1$, let $\begin{bmatrix} t_0 \\ t_1 \end{bmatrix} := x_0$. Then it follows from the split-radix splitting that

$$P_m^T p'_1 = \begin{bmatrix} z_0 \\ J_l \bar{z}_0 \end{bmatrix},$$

where $z_0 = W_l D''_l(t_0 + it_1)$. Note that the computation of z_0 from x_0 requires $(l - 1)\mu_{\mathbb{C}}$, including the one eighth root of unity when $l > 1$. Thus, p is calculated from u_0 in 1(a) and 1(b) using $\tau(m) + \phi(l) + (2l - 2)\alpha_{\mathbb{R}} + (4l - 6)\mu_{\mathbb{R}}$ operations, and likewise for q.

We now show that r and s can be obtained by negating the odd parts of $\bar{p}$ and $\bar{q}$. We have

$$r = W_n E_n \begin{bmatrix} J_m x_0 \\ 0 \end{bmatrix} = W_n E_n \begin{bmatrix} 0 & J_m \\ J_m & 0 \end{bmatrix} \begin{bmatrix} 0 \\ x_0 \end{bmatrix} = W_n E_n J_n \begin{bmatrix} 0 \\ x_0 \end{bmatrix}.$$

It is easily verified that $E_n J_n = K_n$ and $W_n E_n J_n = \bar{W}_n$, so that

$$\bar{r} = W_n \begin{bmatrix} 0 \\ x_0 \end{bmatrix}.$$

Hence,

$$\begin{bmatrix} \bar{r}'_0 \\ \bar{r}'_1 \end{bmatrix} := P_n^T \bar{r} = \begin{bmatrix} W_m & W_m \\ W_m D'_m & -W_m D'_m \end{bmatrix} \begin{bmatrix} 0 \\ x_0 \end{bmatrix},$$

and so

$$\bar{r}'_0 = W_m x_0 = p'_0, \qquad \bar{r}'_1 = -W_m D'_m x_0 = -p'_1.$$

The same relationship holds for s and q. Thus $[r, s]$ can be obtained from $[p, q]$ with no additional computation.

The counts for the computations in steps 1(e) and 3(c) follow from the fact that the transforms are determined by $m - 1$ complex and two real numbers. Also note that, since $\Pi_n p = \begin{bmatrix} \Pi_m p_0 \\ \Pi_m p_1 \end{bmatrix}$, the above manipulations are easily performed when the transforms are stored in bit-reversed order.

The amount of computation $\omega(n)$ required to obtain u, v by the generalized Schur algorithm with real input data therefore satisfies, for $n > n_0 > 2$,

$$\omega(n) = 2\omega(n/2) + 4\tau(n) + 4\tau(n/2) + 4\phi(n/4) + (14n - 24)\alpha_{\mathbb{R}} + (20n - 40)\mu_{\mathbb{R}}$$

$$= 2\omega(n/2) + 16n \lg n - 8n + 16 \quad \text{total real operations,}$$

with roughly twice as many additions as multiplications. By Lemma 2.1, we obtain

$$\omega(n) = 8n \lg^2 n + Cn - 16 \qquad (n \geqq n_0), \tag{4.1}$$

where C is determined by $\omega(n_0)$.

Note that

$$\xi_{0,1} = \gamma_1 = \alpha_{0,0}/\beta_{0,0} \qquad \eta_{0,1} = 1,$$

so that $\omega(1) = 1\mu_{\mathbb{R}}$. If we use the doubling procedure at this stage, we obtain $\omega(2) = 16\alpha_{\mathbb{R}} + 18\mu_{\mathbb{R}} = 34\tau_{\mathbb{R}}$, $\omega(4) = 180$ and $C = 17$. We can improve on this by considering more direct methods to carry out the recursions in the early stages of the algorithm.

Note that we can obtain $\xi_{0,2}$, $\eta_{0,2}$ from $\alpha_0^{(2)}$, $\beta_0^{(2)}$ using $7\tau_{\mathbb{R}}$ from

$$\xi_{0,2} \leftrightarrow x_0 := \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \qquad \eta_{0,2} \leftrightarrow y_0 := \begin{bmatrix} 1 \\ \gamma_1 \gamma_2 \end{bmatrix}$$

where

$$\gamma_1 = \alpha_{0,0}/\beta_{0,0}, \qquad \gamma_2 = \frac{\alpha_{1,0}}{\beta_{1,0}} = \frac{\beta_{0,1}\gamma_1 - \alpha_{0,1}}{\alpha_{0,1}\gamma_1 - \beta_{0,0}}.$$

More generally, k steps of the progressive Schur algorithm can be used to generate the first k Schur parameters in $2k(k+1)\tau_{\mathbb{R}}$. Then ξ_k and η_k can be obtained in $(2k^2 - 5k + 3)\tau_{\mathbb{R}}$ using (3.1), and u, $v \in \mathbb{R}^k$ are obtained in $2\tau(k)$ additional operations. Using this procedure, we obtain $\omega(4) = 67\tau_{\mathbb{R}}$ and $C = -45/4$ in (4.1).

Note that $\omega(n)$ is the amount of computation for u, $v \in \mathbb{C}^n$. An additional $2\tau(n)$ is needed to obtain ξ_n, η_n. We therefore have the following.

PROPOSITION 4.1. *Given the first n terms of formal power series α, β such that α/β is a Schur function, the amount of computation required by the generalized Schur algorithm to obtain the nth Schur polynomials ξ_n, η_n is at most*

$$8n \lg^2 n + 4n \lg n - \frac{77}{4} n - 4$$

real arithmetic operations for $n = 2^\nu > 2$.

5. Application to real Toeplitz systems of equations. The real positive definite Toeplitz matrix $M = M_{n+1} = [\mu_{j-k}]_{j,k=0}^{n} = M^T$ defines an inner product on $\mathbb{R}_{n+1}[\lambda]$ by $\langle \lambda^j, \lambda^k \rangle := \mu_{j-k}$, and the monic polynomials $\chi_k(\lambda)$ that are orthogonal under this inner product are the monic *Szegö polynomials* determined by M. Let $\delta_k^2 := \langle \chi_k, \chi_k \rangle$, and define R to be the unit right triangular matrix whose kth column contains the coefficients of $\chi_k (0 \leqq k \leqq n)$. Then we have

$$R^T M R = D := \operatorname{diag} [\delta_k]_0^n,$$

$$M^{-1} = R D^{-1} R^{-T},$$

so the Szegö polynomials determine the *reverse Choleski factorization* of M^{-1}. These polynomials satisfy the recurrence relations below, which comprise the first phase of many fast Toeplitz solvers, particularly those of Levinson and Trench (see, e.g., [21], [29], [16]). Let $\chi_k \leftrightarrow [r_1^k] \in \mathbb{R}^{k+1}$ and $m_k := [\mu_j]_1^k \in \mathbb{R}^k$.

LEVINSON'S ALGORITHM (Szegö Recursions).

input: $[\mu_j]_0^n$

$\delta_0 := \mu_0$, $\gamma_1 := -\mu_1/\mu_0$,

$r_1 := \gamma_1$, $\delta_1 = (1 - \gamma_1^2)\delta_0$,

for $k = 1, \cdots, n-1$ **do**

$$\gamma_{k+1} = \frac{-1}{\delta_k} m_{k+1}^T \begin{bmatrix} r_k \\ 1 \end{bmatrix},$$

$$r_{k+1} = \begin{bmatrix} 0 \\ r_k \end{bmatrix} + \begin{bmatrix} 1 \\ \tilde{r}_k \end{bmatrix} \gamma_{k+1},$$

$$\delta_{k+1} = \delta_k (1 - \gamma_{k+1}^2).$$

Thus, χ_k, δ_k, $\gamma_k (0 < k \leqq n)$ can be obtained using $n^2 \alpha_{\mathbb{R}} + n(n+2)\mu_{\mathbb{R}}$.

In Levinson's algorithm the χ_k, δ_k are used to solve $Mx = b$ using the inverse Choleski factorization. In Trench's algorithm the nth degree polynomial χ_n and its norm δ_n are used to construct M_n^{-1} by means of the classical Christoffel–Darboux–Szegö formula; the matrix interpretation is the *Gohberg–Semencul formula*:

$$\delta_n M_{n+1}^{-1} = T_1 T_1^T - T_0^T T_0$$

where

$$T_0 = \begin{bmatrix} 0 & & \cdots & 0 \\ \rho_0 & 0 & & \\ \rho_1 & \rho_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \\ \rho_{n-1} & \rho_{n-2} & \cdots & \rho_0 \quad 0 \end{bmatrix}, \qquad T_1^T = \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ \rho_{n-1} & 1 & & & \\ \rho_{n-2} & \rho_{n-1} & \ddots & & \vdots \\ \vdots & \vdots & \ddots & & \\ \rho_0 & \rho_1 & \cdots & \rho_{n-1} & 1 \end{bmatrix}$$

and $\chi_n(\lambda) = \sum_0^n \rho_j \lambda^j$ (see, e.g., [15]).

In contrast with the Szegö recursions, the progressive Schur algorithm is used to obtain the Choleski decomposition of M, $M = LDL^T$ where L is unit left triangular. In particular, this factorization is obtained in $O(n^2)$ operations by performing n steps of Schur's classical algorithm applied to a Schur function $\phi_0 = \alpha_0/\beta_0$ with

$$\alpha_0^{(n)} = -\sum_{j=0}^{n-1} \mu_{j+1} \lambda^j, \qquad \beta_0^{(n)} = \sum_{j=0}^{n-1} \mu_j \lambda^j. \tag{5.1}$$

Moreover, the Schur parameters generated by this process are the same as the Schur parameters generated by the Szegö recursions.

The generalized Schur algorithm with the above initialization does not generate the Choleski factorization, but only pieces of it. Nevertheless, all the Schur parameters are generated, and moreover, it follows from the recursions that the Schur polynomials determine the Szegö polynomials [3]; specifically,

$$\chi_n = \tilde{\eta}_n + \tilde{\xi}_n/\lambda.$$

Thus, we can use the generalized Schur algorithm to obtain ξ_n, η_n, compute χ_n and use the Gohberg–Semencul formula to obtain $M^{-1}b$. In this way the Toeplitz system is solved using $8n \lg^2 n + O(n \lg n)$ total real arithmetic operations.

The operation count for the generalized Schur algorithm can be reduced for this superfast Toeplitz solver. Some observations to this effect follow.

Note that since we are not interested in the Schur polynomials per se, we can get χ_n from the transforms $[u, v] = W_n[x, y]$, saving one FFT of order n. Let $x, y \leftrightarrow \xi_n, \eta_n$. Then

$$\chi_n \leftrightarrow \begin{bmatrix} r_n \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ J_n y_n \end{bmatrix} + \begin{bmatrix} J_n \chi_n \\ 0 \end{bmatrix},$$

and since $\eta_n(0) = 1$,

$$r_n = J_n x + E_n J_n y_n - e_0.$$

It follows that

$$r_n = F_n(\overline{D_n u + v}) - e_0.$$

Note that $D_n u$ has CE symmetry, so this product involves two multiplications by eighth roots of unity when $n > 4$. We can therefore obtain χ_n from u, v using $\tau(n) + (2n - 3)\alpha_{\mathbb{R}} + (2n - 12)\mu_{\mathbb{R}}$ operations. However, multiplication by D_n requires the use of the bit-reversal permutation. In particular, letting $w_n = [\omega_n^j]_0^{n-1} \in \mathbb{C}^n$, we have $\Pi_n D_n u = \Pi_n(w_n \cdot u) = \Pi_n w_n \cdot \Pi_n u$, so we need the nth roots of unity in bit-reversed form. We can either use an additional storage vector for the roots of unity in bit-reversed order, or the correctly ordered w_n can be shuffled when needed. Since this shuffling replaces one FFT, it does not increase the number of data accesses, but it does provide some computational savings.

The relationship between $a \leftrightarrow \alpha_0^{(n)}$ and $b \leftrightarrow \beta_0^{(n)}$ provides more opportunity to reduce the amount of computation if we are content to shuffle data in order to avoid an FFT. We have $b = (\mu_0 - \mu_n)e_0 - E_n a$, and since $W_n E_n = D_n W_n$,

$$d = (\mu_0 - \mu_n)e - D_n c,$$

where $e := W_n e_1$ is the vector of all ones, and $[c, d] := W_n[a, b]$. Thus, when $n \geqq 8$, we can replace $1\tau(n)$ with $(n/2 - 2)\mu_{\mathbb{C}}$ including two multiplications by eighth roots of unity and $(n/2 + 2)\alpha_{\mathbb{R}}$, or $(7/2)n - 14\tau_{\mathbb{R}}$. This results in a savings of $2n \lg n - (15/2)n + 20$ total real operations, but we must have $\Pi_n w_n$ with $\Pi_n c$ to obtain $\Pi_n D_n c$. Clearly, this procedure can be used each time the elements of α_0, β_0 are accessed; however, it cannot be used in the later stages of the algorithm because $\alpha_m, \beta_m (m > 0)$ are not related as α_0, β_0 are. Since $\alpha_0^{(k)}, \beta_0^{(k)}$ are used for each $4 < k = 2^\kappa \leqq n$, the total savings of using this procedure is (by Lemma 2.3)

$$4n \lg n - 19n + 20 \lg n + 4\tau_{\mathbb{R}} \qquad (n > 4).$$

The above observations yield the following.

THEOREM 5.1. *The computation of* χ_n *and* δ_n *from the real positive definite Toeplitz matrix* $M = [\mu_{i-j}] = M^T$ *using the generalized Schur algorithm as described above requires at most*

$$8n \lg^2 n - 2n \lg n + \frac{31}{4} n - 20 \lg n - 29$$

real arithmetic operations for $n = 2^\nu > 4$.

Thus, the total operation count is slightly less than $8n \lg^2 n$, while that of Levinson's algorithm is slightly more than $2n^2$. These bounds are equal at $n = 256$, so the amount of computation for this algorithm is less than that of Levinson's algorithm for $n \geqq 256$. Of course, this difference rapidly becomes substantial as n doubles. We remark that the algorithm uses roughly twice as many additions as multiplications.

6. Concluding remarks. The fact that Levinson's algorithm is more efficient for $n < 256$ indicates some improvement in this superfast Toeplitz solver must be possible in its early stages. Moreover, the split Levinson recursions [11], in which redundancies in Levinson's algorithm are removed, requires about $3n^2/2$ real operations. This indicates the likelihood of improving the implementation of the generalized Schur algorithm for the superfast solution of Toeplitz systems. In fact, we made little use of the relationship between α_0 and β_0 given by (5.1).

While the algorithm may vectorize well, the above description is inherently sequential because the computation of $\xi_{m,m}$ cannot proceed until $\alpha_m^{(m)}$ and $\beta_m^{(m)}$ are calculated, which in turn cannot be computed until $\xi_{0,m}$ and $\eta_{0,m}$ are computed. These bottlenecks in the doubling strategy show how this algorithm is not a splitting into independent subproblems as in the case of an FFT. Thus, apart for the obvious independent quantities to be calculated within a step (e.g., p, q can be calculated from u_0, v_0 simultaneously), any parallel implementation of the algorithm is likely to be inherently different from the one presented here.

With regard to the use of the generalized Schur algorithm in the case that n is not a power of two, the recursions (3.2), (3.3) can be used to derive the appropriate convolution formulas for a given factorization of n. For example, if $n = 3m$ a decomposition of ξ_n, η_n into three smaller Schur polynomials that correspond with Schur functions ϕ_0, ϕ_m, ϕ_{2m} could be derived. In this manner the development of multiple radix implementations of the generalized Schur algorithm may proceed analogously with that of FFT algorithms. In a sense, the generalized Schur algorithm is one level of complexity higher than an FFT. It is not unlikely that a family of implementations of the generalized Schur algorithm will be useful in practice, each one tailored to specific lengths of and symmetries in the input data.

We finally remark that we have strived for the lowest operation count for aesthetic and theoretical rather than practical reasons. Some of the fine tuning described in § 5 will not appear in code for the algorithm because it will make the code too long and tedious. We will, nevertheless, use the dual split-radix codes to reduce the amount of computation and avoid the bit-reversal permutation.

Acknowledgment. We are indebted to Avideh Zakhor for helpful discussions on the split-radix FFT algorithms.

REFERENCES

[1] N. I. AKHIEZER, *The Classical Moment Problem*, Oliver and Boyd, Edinburgh, 1965.

[2] G. S. AMMAR AND W. B. GRAGG, *Implementation and use of the generalized Schur algorithm*, in Com-

putational and Combinatorial Methods in Systems Theory, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 265–280.

[3] ———, *The generalized Schur algorithm for the superfast solution of Toeplitz systems*, in Rational Approximation and its Applications in Mathematics and Physics, J. Gilewicz, M. Pindor, and W. Siemaszko, eds., Lecture Notes in Mathematics 1237, Springer-Verlag, New York, Berlin, 1987, pp. 315–330.

[4] E. H. BAREISS, *Numerical solution of linear equations with Toeplitz and vector Toeplitz matrices*, Numer. Math., 13 (1969), pp. 404–424.

[5] G. D. BERGLAND, *A fast Fourier transform algorithm for real-valued series*, Comm. ACM, 11 (1968), pp. 703–710.

[6] R. R. BITMEAD AND B. D. O. ANDERSON, *Asymptotically fast solution of Toeplitz and related systems of linear equations*, Linear Algebra Appl., 34 (1980), pp. 103–116.

[7] R. E. BLAHUT, *Fast Algorithms for Digital Signal Processing*, Addison-Wesley, Reading, MA, 1985.

[8] R. P. BRENT, F. G. GUSTAVSON, AND D. Y. Y. YUN, *Fast solution of Toeplitz systems of equations and computation of Padé approximants*, J. Algorithms, 1 (1980), pp. 259–295.

[9] W. L. BRIGGS, *Further symmetries of in-place* FFTs, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 644–654.

[10] O. BUNEMAN, *Conversion of* FFTs *to fast Hartley transforms*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 624–638.

[11] P. DELSARTE AND Y. V. GENIN, *The Split Levinson Algorithm*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 470–478.

[12] F. DE HOOG, *A new algorithm for solving Toeplitz systems of equations*, Linear Algebra Appl., 88/89 (1987), pp. 122–138.

[13] P. DUHAMEL, *Implementation of "split-radix"* FFT *algorithms for complex, real, and real-symmetric data*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 285–295.

[14] L. Y. GERONIMUS, *Orthogonal Polynomials*, Consultants Bureau, New York, 1961.

[15] I. C. GOHBERG AND I. A. FEL'DMAN, *Convolution Equations and Projection Methods for their Solution*, American Mathematical Society, Providence, RI, 1974.

[16] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1984.

[17] J. R. JAIN, *An efficient algorithm for a large Toeplitz set of linear equations*, IEEE Trans. Acoust. Speech Signal Process., 27 (1979), pp. 612–615.

[18] T. KAILATH, *A theorem of I. Schur and its impact on modern signal processing*, in I. Schur Methods in Operator Theory and Signal Processing, I. C. Gohberg, ed., Birkhäuser-Verlag, Basel, 1986, pp. 9–30.

[19] T. KAILATH, A. VIEIRA, AND M. MORF, *Inverses of Toeplitz operators, innovations and orthogonal polynomials*, SIAM Rev., 20 (1978), pp. 106–119.

[20] R. KUMAR, *A fast algorithm for solving a Toeplitz system of equations*, IEEE Trans. Acoust. Speech Signal Process., 33 (1985), pp. 254–267.

[21] N. LEVINSON, *The Wiener* RMS *(Root-Mean-Square) error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[22] B. R. MUSICUS, *Levinson and fast Choleski algorithms for Toeplitz and almost Toeplitz matrices*, Report, Research Lab. of Electronics, Massachusetts Institutes of Technology, Cambridge, MA, 1984.

[23] I. SCHUR, *Uber Potenzreihen, die in Innern des Einheitskrises Beschränkt Sind*, J. Reine Angew. Math., 147 (1917), pp. 205–232.

[24] H. V. SORENSEN, M. T. HEIDEMAN, AND C. S. BURRUS, *On computing the split-radix* FFT, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 152–156.

[25] H. V. SORENSEN, D. L. JONES, M. T. HEIDEMAN, AND C. S. BURRUS, *Real-valued Fast Fourier Transform algorithms*, IEEE Trans. Acoust. Speech Signal Process., 35 (1987), pp. 849–863.

[26] J. STOER AND R. BULIRSCH, *Introduction to Numerical Analysis*, Springer-Verlag, New York, Berlin, 1980.

[27] P. N. SWARZTRAUBER, *Symmetric* FFTs, Math. Comp. 47 (1986), pp. 323–346.

[28] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1939.

[29] W. 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. 9, No. 1, January 1988

006

A PENCIL APPROACH FOR EMBEDDING A POLYNOMIAL MATRIX INTO A UNIMODULAR MATRIX*

T. BEELEN† AND P. VAN DOOREN‡

Abstract. In this paper a new method for constructing the unimodular embedding of a polynomial matrix $P(\lambda)$ is derived. As proposed by Eising, the problem can be transformed to one of embedding a pencil, derived from the polynomial matrix $P(\lambda)$. The actual embedding of the pencil is performed here via the staircase form of this pencil, which shortcuts Eising's construction. This then leads to a *new, fast, and numerically reliable* algorithm for embedding a polynomial matrix. The new method uses a fast variant of the staircase algorithm and only requires $O(p^3)$ operations in contrast to the $O(p^4)$ methods proposed up to now (where p is the largest dimension of the pencil). At the same time we also treat the connected problem of finding the (right) null space and (right) inverse of a polynomial matrix $P(\lambda)$.

Key words. polynomial matrix, unimodular matrix, staircase form

AMS(MOS) subject classifications. 65F30, 93B10

1. Introduction. Let $P(\lambda)$ be an $m \times n$ (with $m < n$) polynomial matrix of degree d:

$$P(\lambda) \doteq P_0 + P_1\lambda + P_2\lambda^2 + \cdots + P_d\lambda^d \tag{1}$$

where each P_i is a real or complex $m \times n$ matrix. In this paper we develop a new algorithm to construct an embedding of this polynomial matrix into a unimodular one, i.e., to find a second polynomial matrix $Q(\lambda)$ of dimension $(n - m) \times n$:

$$Q(\lambda) \doteq Q_0 + Q_1\lambda + \cdots + Q_{d_q}\lambda^{d_q} \tag{2}$$

such that the compound matrix

$$U(\lambda) \doteq \begin{bmatrix} P(\lambda) \\ Q(\lambda) \end{bmatrix} \tag{3}$$

is unimodular.

Since a unimodular matrix is by definition invertible for all $\lambda \in \mathbf{C}$ (where $\mathbf{C}$ is the *finite* complex plane), the submatrix $P(\lambda)$ must necessarily have full row rank m for all $\lambda \in \mathbf{C}$ in order for a solution of the embedding problem to exist. It turns out that this is also a sufficient condition for a solution to exist and that, moreover, there always exists a solution $Q(\lambda)$ of degree $d_q \leqq d - 1$ [4]. (Here we assumed that $d \geqq 1$ since otherwise the problem degenerates into one involving constant matrices only and becomes trivial.) Although this result was known it is nice to see how easily it is also derived from our algorithmic construction.

The constructive method developed in this paper is then shown to be easily extended to one that also provides the right inverse of $P(\lambda)$, i.e., an $n \times m$ polynomial matrix $M(\lambda)$ satisfying

$$P(\lambda) \cdot M(\lambda) = I_m \tag{4}$$

* Received by the editors March 1, 1987; accepted for publication October 1, 1987. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† PICOS—Glass, Philips, Eindhoven, the Netherlands.

‡ Philips Research Laboratory, Brussels, Belgium.

and the right null space of $P(\lambda)$, i.e., an $n \times (n - m)$ polynomial matrix $N(\lambda)$ of full column rank and satisfying

$$P(\lambda) \cdot N(\lambda) = 0_{m,n-m}. \tag{5}$$

Our method reformulates the embedding problem of $P(\lambda)$ as an embedding of a *pencil*, an idea which was, e.g., used by Eising [2]. After recalling this in § 2, we show in the next section how the embedding problem for pencils can be trivially solved via the staircase algorithm [10] of pencils. In § 4, we then use these ideas to provide algorithms for solving the related equations for the right inverse and right null space of $P(\lambda)$. Finally we conclude in § 5 with some considerations of complexity and numerical stability of our method and with some numerical examples.

2. Reduction to a pencil problem. The idea developed here is borrowed from Eising [2]. Consider the $dm \times \{(d-1)m + n\}$ pencil $\lambda B - A$ where the matrices B and A are defined as

$$B \doteq \begin{bmatrix} 0 & & -P_d \\ I_m & \ddots & -P_{d-1} \\ & \ddots & \vdots \\ & & -P_2 \\ & I_m & -P_1 \end{bmatrix}, \qquad A \doteq \begin{bmatrix} I_m & & & \\ & \ddots & & \\ & & I_m & \\ & & & P_0 \end{bmatrix}. \tag{6}$$

We first show that the pencil $\lambda B - A$ has full row rank for all $\lambda \in \mathbf{C}$ if and only if the polynomial matrix $P(\lambda)$ has full row rank for all $\lambda \in \mathbf{C}$. For this we introduce the $dm \times dm$ unimodular matrices $C(\lambda)$ and $D(\lambda) = C^{-1}(\lambda)$ defined as

$$C(\lambda) \doteq \begin{bmatrix} I & & & \\ \lambda I & & & \\ \lambda^2 I & & & \\ \vdots & & \ddots & \\ \lambda^{d-1} I & \cdots & \lambda^2 I & \lambda I & I \end{bmatrix}, \qquad D(\lambda) \doteq \begin{bmatrix} I & & & \\ -\lambda I & & & \\ & \ddots & & \\ & & -\lambda I & I \end{bmatrix} \tag{7}$$

where all identity matrices are of order m. Indeed, by straightforward calculations we find

$$C(\lambda)(A - \lambda B) = C(\lambda) \begin{bmatrix} I & & & \lambda P_d \\ -\lambda I & \ddots & & \vdots \\ & \ddots & I & \lambda P_2 \\ & & -\lambda I & \lambda P_1 + P_0 \end{bmatrix} = \begin{bmatrix} I & & & R_d(\lambda) \\ & \ddots & & \vdots \\ & & I & R_2(\lambda) \\ & & & P(\lambda) \end{bmatrix} \tag{8}$$

where we define $R_{d+1}(\lambda) \doteq 0$, $R_i(\lambda) \doteq \lambda R_{i+1}(\lambda) + \lambda P_i$, $i = d, \cdots, 2$ and $R_1(\lambda) \doteq P(\lambda)$. Using this, and the fact that $C(\lambda)$ is unimodular (and hence invertible for all $\lambda \in \mathbf{C}$) we indeed easily derive that the pencil $\lambda B - A$ has full row rank for all $\lambda \in \mathbf{C}$ if and only if $P(\lambda)$ has full row rank for all $\lambda \in \mathbf{C}$.

Suppose now that we are able to provide an embedding for the pencil $\lambda B - A$, which we denote as

$$\begin{bmatrix} \lambda B - A \\ K(\lambda) \end{bmatrix} \tag{9}$$

and let us partition $K(\lambda)$ as follows:

$$K(\lambda) \doteq [K_1(\lambda), \cdots, K_{d-1}(\lambda), K_d(\lambda)] \tag{10}$$

where $K_i(\lambda)$ has dimensions $(n-m)\times m$, for $i<d$ and $K_d(\lambda)$ dimensions $(n-m)\times n$. Combining (8) and (9), we thus have that

$$G(\lambda)\doteq\left[\begin{array}{ccc|c} I_m & & & R_d(\lambda) \\ & \ddots & & \vdots \\ & & I_m & R_2(\lambda) \\ 0 & \cdots & 0 & P(\lambda) \\ \hline K_1 & \cdots & K_{d-1} & K_d \end{array}\right]=\left[\begin{array}{c|c} -C(\lambda) & \\ \hline & I_{n-m} \end{array}\right]\cdot\left[\begin{array}{c} \lambda B-A \\ \hline K(\lambda) \end{array}\right] \tag{11}$$

is also unimodular. Introducing the unimodular matrix $H(\lambda)$ of order $(d-1)m+n$ as

$$H(\lambda)\doteq\left[\begin{array}{ccc|c} I_m & & & \\ & \ddots & & \\ & & I_m & \\ & & & I_m \\ \hline -K_1 & \cdots & -K_{d-1} & 0 \end{array}\begin{array}{|c} \\ \\ \\ \\ \hline I_{n-m} \end{array}\right] \tag{12}$$

and premultiplying $G(\lambda)$ by $H(\lambda)$ gives

$$S(\lambda)\doteq H(\lambda)G(\lambda)=\left[\begin{array}{ccc|c} I_m & & & R_d(\lambda) \\ & \ddots & & \vdots \\ & & I_m & R_2(\lambda) \\ \hline & & & P(\lambda) \\ & & & Q(\lambda) \end{array}\right] \tag{13}$$

where $Q(\lambda)$ is given by

$$Q(\lambda)\doteq K_d(\lambda)-\sum_{i=1}^{d-1}K_i(\lambda)R_{d-i+1}(\lambda). \tag{14}$$

It is now obvious from (13) that

$$\left[\begin{array}{c} P(\lambda) \\ Q(\lambda) \end{array}\right] \tag{15}$$

is unimodular if and only if the embedding (9) is unimodular. This thus shows that the problem of embedding a polynomial matrix (provided this is possible) can always be reformulated as that of embedding a pencil. The reason of reformulating the problem as one for a pencil is that it can be embedded by a *constant* matrix K, as was, e.g., shown by Eising [2]. In the next section we give a simple alternative proof of this result and also show how to construct such a *constant* solution K.

3. Embedding a pencil in a unimodular one. Kronecker (see [3]) has shown that any pencil $\lambda B-A$ can be transformed via *constant* invertible column and row transformations to a canonical block diagonal form λB_c-A_c

$$S\cdot(\lambda B-A)\cdot T=\lambda B_c-A_c=\operatorname{diag}\{L_{\varepsilon_1},\cdots,L_{\varepsilon_p},L_{\eta_1}^T,\cdots,L_{\eta_q}^T,\lambda N-I,\lambda I-J\} \tag{16}$$

where

(1) L_ε is the $\varepsilon\times(\varepsilon+1)$ bidiagonal pencil

$$\left[\begin{array}{ccccc} -1 & \lambda & & & \\ & \ddots & \ddots & & \\ & & & -1 & \lambda \end{array}\right]. \tag{17}$$

(2) L_η^T is the $(\eta + 1) \times \eta$ bidiagonal pencil

$$\begin{bmatrix} -1 & & \\ \lambda & \ddots & \\ & \ddots & -1 \\ & & \lambda \end{bmatrix}. \tag{18}$$

(3) N is a nilpotent Jordan matrix, and hence $\lambda N - I$ consists of a diagonal block pencil with $\delta_i \times \delta_i$ blocks of the type

$$\begin{bmatrix} -1 & \lambda & & \\ & -1 & \ddots & \\ & & \ddots & \lambda \\ & & & -1 \end{bmatrix}. \tag{19}$$

(4) J is in Jordan canonical form.

The matrix $\lambda I - J$ contains the *finite elementary divisors* and $\lambda N - I$ the *infinite elementary divisors* of $\lambda B - A$. The blocks L_{ε_i} and $L_{\eta_j}^T$ contain the *singularity* of the pencil. The indices ε_i and η_j are called the *Kronecker column* and *row indices*, respectively, and δ_i are called the *degrees* of the infinite elementary divisors.

Using this canonical form we now easily derive the following theorem about the unimodular embedding of a pencil.

THEOREM 1. *A pencil $\lambda B - A$ has a unimodular embedding*

$$\begin{bmatrix} \lambda B - A \\ K(\lambda) \end{bmatrix} \tag{20}$$

if and only if it has no finite elementary divisors and no Kronecker row indices. Moreover, there always exists a constant matrix K such that the new infinite elementary divisors of the embedding are equal to the union of the infinite elementary divisors $\{\delta_i\}$ and of the Kronecker column indices $\{(\varepsilon_j + 1)\}$ of $\lambda B - A$.

Proof. The necessity of the condition is trivial as noted in the Introduction. Indeed the unimodular embedding has full (row) rank for all $\lambda \in \mathbf{C}$, and thus this is also implied for the rows of $\lambda B - A$. Using the block decomposition (16) we easily find then that $\lambda B - A$ can have no finite elementary divisors or Kronecker row indices, since the corresponding blocks do not obey the row rank property for all $\lambda \in \mathbf{C}$.

The sufficiency of the condition is now proved via the construction of a solution K, which at the same time satisfies the second part of the theorem. Indeed, choose K_c to be a matrix whose rows are unit vectors, each with a -1 at the location corresponding to the last column of one of the L_{ε_j} of $\lambda B_c - A_c$. Then obviously the embedding

$$\begin{bmatrix} \lambda B_c - A_c \\ K_c \end{bmatrix} \tag{21}$$

has a Kronecker canonical form with blocks (19) of sizes δ_i and $(\varepsilon_j + 1)$ as requested. This form is indeed obtained by a mere permutation of the rows of (21). Then, defining $K \doteq K_c \cdot T^{-1}$ and using

$$\left[\begin{array}{c|c} S & \\ \hline & I \end{array}\right] \cdot \left[\begin{array}{c} \lambda B - A \\ \hline K \end{array}\right] \cdot T = \left[\begin{array}{c} \lambda B_c - A_c \\ \hline K_c \end{array}\right] \tag{22}$$

we find that (20) and (21) have the same Kronecker canonical form. The fact that a pencil with only infinite elementary divisors is unimodular [4] then completes the proof. □

COROLLARY 1. *A polynomial matrix $P(\lambda)$ of degree d has a unimodular embedding*

$$\begin{bmatrix} P(\lambda) \\ Q(\lambda) \end{bmatrix} \tag{23}$$

if and only if it has full row rank for all $\lambda \in \mathbf{C}$. Moreover, there always exists an embedding with a polynomial matrix $Q(\lambda)$ of degree $(d-1)$.

Proof. As above, the necessity of the condition is trivial. Sufficiency is proved via the construction of K above and the subsequent derivation of $Q(\lambda)$ in (9)–(14). If K is chosen constant, then the construction (14) and the recurrence relation for $R_i(\lambda)$ in (8) yields the following explicit formula for $Q(\lambda)$ in terms of the coefficient matrices P_i of $P(\lambda)$:

$$Q(\lambda) \doteq K_d - \sum_{i=1}^{d-1} K_i \sum_{j=0}^{i-1} \lambda^{i-j} P_{d-j} = K_d - \sum_{k=1}^{d-1} \lambda^k \sum_{i=k}^{d-1} K_i P_{d+k-i}. \tag{24}$$

This clearly shows that $Q(\lambda)$ has degree $(d-1)$ and thus completes the proof. □

While apparently the problem is thus solved via the above construction, it is not a recommended procedure from a numerical point of view. The transformations S and T in the decomposition (16) may indeed be very badly conditioned and thus give rise to a significant loss of accuracy. An alternative decomposition that does not suffer from this drawback is the so-called staircase form of $\lambda B - A$ [10]. For a pencil $\lambda B - A$ with *only* column Kronecker indices $\{\varepsilon_j\}$ and infinite elementary divisors $\{\delta_j\}$, we obtain the following staircase form (which we denote by $\lambda B_{e\infty} - A_{e\infty}$) via *unitary* transformations U and V [10]:

$$U(\lambda B - A)V = \lambda B_{e\infty} - A_{e\infty}$$

$$= \begin{bmatrix} -A_{1,1} & \lambda B_{1,2} - A_{1,2} & X & \cdots & X \\ & -A_{2,2} & \lambda B_{2,3} - A_{2,3} & & \vdots \\ & & \ddots & \ddots & X \\ & & & -A_{k,k} & \lambda B_{k,k+1} - A_{k,k+1} \\ & & & & -A_{k+1,k+1} \end{bmatrix}. \tag{25}$$

This form is characterized by the fact that the blocks $A_{i,i}$ $(i = 1, \cdots, k+1)$ have full row rank and the blocks $B_{i,i+1}$ $(i = 1, \cdots, k)$ have full column rank. Notice that the blocks indicated by X in (25) are in fact pencils as well. Let the matrices $A_{i,i}$ $(i = 1, \cdots, k+1)$ and $\lambda B_{i,i+1} - A_{i,i+1}$ $(i = 1, \cdots, k)$ have dimensions $m_i \times n_i$ $(m_i \leqq n_i)$ and $m_i \times n_{i+1}$ $(n_{i+1} \leqq m_i)$, respectively. Then the following theorem, proved in [10], relates these dimensions to the Kronecker canonical form of $\lambda B_{e\infty} - A_{e\infty}$ (or $\lambda B - A$).

THEOREM 2. *The pencil $\lambda B - A$ with staircase form as in* (25), *has*

$$\begin{aligned} &n_i - m_i && \text{Kronecker column indices } \varepsilon_j \text{ equal to } i-1, \\ &m_i - n_{i+1} && \text{infinite elementary divisors } \delta_j \text{ equal to } i. \end{aligned} \tag{26}$$

□

At first sight we thus have the requested information to find a constant matrix K for the embedding, using the decomposition (25) as well. That this is in fact very simple is now shown below. Corresponding to each nonsquare $A_{i,i}$ we can easily find and $(n_i - m_i) \times n_i$ matrix C_i such that

$$\begin{bmatrix} C_i \\ A_{i,i} \end{bmatrix} \tag{27}$$

is square invertible and does not depend on λ. Thus, by adding a block row of the type

$$[0 \cdots 0 \quad -C_i \quad X \cdots X] \tag{28}$$

to each corresponding block row

$$[0 \cdots 0 \quad -A_{i,i} \quad \lambda B_{i,i+1} - A_{i,i+1} \quad X \cdots X] \tag{29}$$

in (25) for ($i = 1, \cdots, k$) and adding $[0 \cdots 0 \quad -C_{k+1}]$ to $[0 \cdots 0 \quad -A_{k+1,k+1}]$, the pencil (25) can be embedded into a pencil

$$\begin{bmatrix} \lambda B_{e\infty} - A_{e\infty} \\ K_{e\infty} \end{bmatrix} \tag{30}$$

with

$$K_{e\infty} \doteq \begin{bmatrix} -C_1 & X & & \cdots & X \\ & -C_2 & \ddots & & \vdots \\ & & & -C_k & X \\ & & & & -C_{k+1} \end{bmatrix}. \tag{31}$$

This matrix $K_{e\infty}$ has dimensions

$$\sum_{i=1}^{k+1} (n_i - m_i) \times \sum_{i=1}^{k+1} n_i = (n-m) \times \{(d-1)m + n\}.$$

It should be noted that the blocks indicated by X in (31) can be chosen arbitrarily, even as a function of λ, so that the matrix (13) is highly nonunique. For the sequel we assume $K_{e\infty}$ to be chosen *constant*. It is easily seen that the pencil (30), up to a row permutation Π_r, is again in staircase form:

$$\Pi_r \cdot \begin{bmatrix} \lambda B_{e\infty} - A_{e\infty} \\ K_{e\infty} \end{bmatrix}$$

$$= \begin{bmatrix} -\begin{bmatrix} C_1 \\ A_{1,1} \end{bmatrix} & \begin{bmatrix} X \\ \lambda B_{1,2} - A_{1,2} \end{bmatrix} & & \cdots & X \\ & -\begin{bmatrix} C_2 \\ A_{2,2} \end{bmatrix} & & & \vdots \\ & & & \ddots & \\ & & \ddots & -\begin{bmatrix} C_k \\ A_{k,k} \end{bmatrix} & \begin{bmatrix} X \\ \lambda B_{k,k+1} - A_{k,k+1} \end{bmatrix} \\ & & & & -\begin{bmatrix} C_{k+1} \\ A_{k+1,k+1} \end{bmatrix} \end{bmatrix} \tag{32}$$

since the (new) blocks $\begin{bmatrix} C_i \\ A_{i,i} \end{bmatrix}$ have full row rank by construction, and the (new) blocks $\begin{bmatrix} 0 \\ B_{i,i+1} \end{bmatrix}$ still have full column rank. The fact that this "embedded" pencil is unimodular easily follows from the full rank property for all $\lambda \in \mathbf{C}$ (guaranteed by the diagonal blocks). The preserved staircase form shows moreover that the pencil (32) has

$$n_i - n_{i+1} \quad \text{infinite elementary divisors } \hat{\delta}_j \text{ equal to } i \tag{33}$$

which, according to Theorem 2, is exactly the same result as in Theorem 1. Just as in (22), we then define $K \doteq K_{e\infty} V^{-1} = K_{e\infty} V^*$ (.* denotes the conjugate transpose) and use

$$\left[\begin{array}{c|c} U & \\ \hline & I \end{array}\right] \cdot \left[\begin{array}{c} \lambda B - A \\ \hline K \end{array}\right] \cdot V = \left[\begin{array}{c} \lambda B_{e\infty} - A_{e\infty} \\ \hline K_{e\infty} \end{array}\right] \tag{34}$$

to show that the matrix K obtained via this construction also satisfies the conditions of Theorem 1. This construction thus implicitly provides an embedding satisfying Theorem 1, *without* passing via the numerically sensitive Kronecker canonical form.

Remark 3.1. The staircase algorithm described for general pencils in [10] in fact also tests whether or not a given pencil only possesses infinite elementary divisors and Kronecker column indices. Applied to the pencil (6) it thus tests for the existence of an embedding, and at the same time provides a convenient form for constructing such an embedding in case it exists.

Remark 3.2. While the general staircase algorithm, e.g., described in [10] or [9] has an operation count that is *quartic* in the maximal dimension dm of $\lambda B - A$ (i.e., $O(m^4d^4)$ flops), an improved method has recently been proposed in [1] which has an operation count that is only *cubic* (i.e., $O(m^3d^3)$ flops). Moreover, it is shown there that the "rank carrying stairs" $A_{i,i}$ and $B_{i,i+1}$ can be chosen *triangular* when appropriately updating U and V.

Remark 3.3. It is well known that in general there is no unique solution $Q(\lambda)$ to the embedding problem. The method described above also does not yield a unique solution $Q(\lambda)$. This is clearly reflected by the freedom in choosing the block rows in (28). A possible selection criterion could be to minimize the effort for determining matrix K. When the $m_i \times n_i$ matrices $A_{i,i}$ in (25) are assumed to be upper triangular the $(n_i - m_i) \times n_i$ matrices C_i $(i \leqq k+1)$ can be chosen as

$$C_i = [I, 0] \tag{35}$$

with the remaining X matrices in each row of $K_{\epsilon\infty}$ equal to 0. In that case, the determination of K is of course trivial.

To conclude this section we now summarize the computational procedure.

ALGORITHM EMBED.

(1) Construct the pencil $\lambda B - A$ defined by (6).
(2) Compute the staircase form of $\lambda B - A$ giving (25) with upper triangular matrices $A_{i,i}$.
(3) Construct matrices C_i satisfying (27).
(4) Compute matrix $K_{\epsilon\infty}$ given in (31).
(5) Determine matrix $Q(\lambda)$ via (34) and (24).

4. Inversion of a unimodular matrix. In this section we consider the problem of inversion of a unimodular matrix from a numerical point of view. Throughout this section we denote by $U(\lambda)$ an $n \times n$ polynomial matrix of degree $d \geqq 1$ that is assumed to be *unimodular*, i.e., such that

$$\det U(\lambda) = \text{a nonzero constant.} \tag{36}$$

The determination of $U^{-1}(\lambda)$ is an important step in several problems dealing with polynomial matrices. For example, this inversion problem arises when solving certain polynomial matrix equations which we now first describe.

Computing a right inverse and a right null space of a full row rank polynomial matrix. Let $P(\lambda)$ denote an $m \times n$ polynomial matrix $(m < n)$ which has full row rank for all $\lambda \in \mathbf{C}$. Any polynomial matrix $M(\lambda)$ such that $P(\lambda)M(\lambda) = I_m$ is called a right inverse of $P(\lambda)$. Any polynomial matrix $N(\lambda)$ of full column rank (for *some* λ) and such

that $P(\lambda)N(\lambda) = 0_{m,n-m}$ is said to span the right null space of $P(\lambda)$. In order to find such matrices $M(\lambda)$ and $N(\lambda)$, we start with *any* unimodular embedding of $P(\lambda)$

$$U(\lambda) \doteq \begin{bmatrix} P(\lambda) \\ Q(\lambda) \end{bmatrix}. \tag{37}$$

This is done using the procedure described in the previous section. Hereafter we determine the inverse of $U(\lambda)$. (To this end, we present a new numerical method in this section.) It is well known (see [4, Lemma 6.3-1]) that this inverse is a polynomial matrix $V(\lambda)$ which we partition as $V(\lambda) = [M(\lambda)|N(\lambda)]$ where $M(\lambda)$ and $N(\lambda)$ have dimensions $n \times m$ and $n \times (n - m)$, respectively. Obviously, we have $U(\lambda)V(\lambda) = I_n$, or equivalently,

$$\left[\frac{P(\lambda)}{Q(\lambda)}\right][M(\lambda)|N(\lambda)] = \left[\begin{array}{c|c} I_m & 0 \\ \hline 0 & I_{n-m} \end{array}\right]. \tag{38}$$

Hence,

$$P(\lambda)M(\lambda) = I_m, \qquad P(\lambda)N(\lambda) = 0_{m,n-m}. \tag{39}$$

Clearly, $M(\lambda)$ is a right inverse of $P(\lambda)$ and $N(\lambda)$ spans its right null space since $N(\lambda)$ has full column rank for all $\lambda \in \mathbf{C}$ (being a submatrix of the unimodular matrix $V(\lambda)$).

From an algebraic point of view the computation of $U^{-1}(\lambda)$ is rather simple. Indeed, let $V(\lambda) = U^{-1}(\lambda)$; then we have to solve

$$U(\lambda)V(\lambda) = I_n. \tag{40}$$

Matrix $U(\lambda)$ can, e.g., be reduced by elementary row operations to the so-called triangular Hermite form (see [4, § 6.3] for details). This form can now be used to solve for $V(\lambda)$ by backward substitution. Of course, other methods for determining $V(\lambda)$ can be applied including those for inverting arbitrary polynomial or rational matrices (see, e.g., [8]). However, most of these general inversion methods are not recommended from a numerical point of view. The main reason is that in fact they rely on the Euclidean division algorithm (when reducing the Hermite form) or on formulas that can cause severe loss of significant digits.

Below we present a new (numerically more reliable) algorithm for computing the inverse of a unimodular matrix. Let us denote the $n \times n$ unimodular matrix $U(\lambda)$ of degree d by

$$U(\lambda) \doteq U_0 + U_1\lambda + U_2\lambda^2 + \cdots + U_d\lambda^d \tag{41}$$

where each U_i is a real or complex $n \times n$ matrix. Here again we assume that $d \geqq 1$, since otherwise the problem degenerates into one involving constant matrices only and becomes trivial.

As in the previous section we reformulate the problem as a pencil problem by defining the $dn \times dn$ pencil $\lambda\hat{B} - \hat{A}$ where the matrices $\hat{B}$ and $\hat{A}$ are defined as

$$\hat{B} \doteq \begin{bmatrix} 0 & & & -U_d \\ I_n & \ddots & & -U_{d-1} \\ & \ddots & 0 & \vdots \\ & & I_n & -U_2 \\ & & & -U_1 \end{bmatrix}, \qquad \hat{A} \doteq \begin{bmatrix} I_n & & & \\ & \ddots & & \\ & & I_n & \\ & & & U_0 \end{bmatrix}. \tag{42}$$

When defining $C(\lambda)$ and $D(\lambda)$ as in (7) but now with identity matrices of order n, we find

$$C(\lambda)(\hat{A}-\lambda\hat{B}) = \begin{bmatrix} I & & \hat{R}_d(\lambda) \\ & \ddots & \vdots \\ & I & \hat{R}_2(\lambda) \\ & & U(\lambda) \end{bmatrix} \tag{43}$$

where we define $\hat{R}_{d+1}(\lambda) \doteq 0$, $\hat{R}_i(\lambda) \doteq \lambda\,\hat{R}_{i+1}(\lambda) + \lambda U_i$, $i = d, \cdots, 2$ and $\hat{R}_1(\lambda) \doteq U(\lambda)$.

It is easily seen that (43) is unimodular. Hence, the pencil $\lambda\hat{B} - \hat{A}$ is also unimodular. It follows from (43) that

$$\begin{aligned} U^{-1}(\lambda) &= -[0, \cdots, 0, I_n](\lambda\hat{B}-\hat{A})^{-1}C^{-1}(\lambda)[0, \cdots, 0, I_n]^T \\ &= -[0, \cdots, 0, I_n](\lambda\hat{B}-\hat{A})^{-1}D(\lambda)[0, \cdots, 0, I_n]^T \\ &= -[0, \cdots, 0, I_n](\lambda\hat{B}-\hat{A})^{-1}[0, \cdots, 0, I_n]^T. \end{aligned} \tag{44}$$

This thus shows that inverting a unimodular polynomial matrix is easily reformulated as inverting a unimodular pencil.

In order now to solve the inversion problem of the unimodular pencil $\lambda\hat{B} - \hat{A}$, we first note that the Kronecker canonical form of $\lambda\hat{B} - \hat{A}$ merely consists of $I - \lambda\hat{N}$ where $\hat{N}$ is nilpotent:

$$\hat{S}\cdot(\lambda\hat{B}-\hat{A})\cdot\hat{T} = I - \lambda\hat{N}. \tag{45}$$

From this the inverse is trivially obtained as

$$(\lambda\hat{B}-\hat{A})^{-1} = \hat{T}^{-1}\cdot(I + \lambda\hat{N} + \lambda^2\hat{N}^2 + \cdots + \lambda^l\hat{N}^l)\cdot\hat{S}^{-1} \tag{46}$$

where $l + 1$ is the size of the largest infinite elementary divisor in (16) (i.e., the largest $\delta_i \times \delta_i$ block of the type (19) in $I - \lambda\hat{N}$). If we define the polynomial matrix $V(\lambda) \doteq U(\lambda)^{-1}$ as

$$V(\lambda) \doteq V_0 + V_1\lambda + \cdots + V_l\lambda^l, \tag{47}$$

then the combination of (44) and (46) gives us

$$V_i = -[0, \cdots, 0, I_n]\cdot\hat{T}^{-1}\cdot\hat{N}^i\cdot\hat{S}^{-1}\cdot[0, \cdots, 0, I_n]^T \qquad (i = 0, \cdots, l) \tag{48}$$

which thus solves the problem. But since the Kronecker decomposition is a sensitive tool from a numerical point of view, we again turn to the staircase form of $\lambda\hat{B} - \hat{A}$. This can be obtained under *unitary transformations* Q and Z:

$$Q\cdot(\lambda\hat{B}-\hat{A})\cdot Z \doteq \lambda\hat{B}_\infty - \hat{A}_\infty$$

$$= \left[\begin{array}{c|c|c|c} -\hat{A}_{1,1} & \lambda\hat{B}_{1,2}-\hat{A}_{1,2} & \cdots & X \\ \hline & -\hat{A}_{2,2} & & \vdots \\ \hline & & \ddots & \lambda\hat{B}_{l,l+1}-\hat{A}_{l,l+1} \\ \hline & & & -\hat{A}_{l+1,l+1} \end{array}\right] \begin{matrix} \} n_1 \\ \\ \} n_l \\ \} n_{l+1} \end{matrix} \tag{49}$$

$$\begin{matrix} \underbrace{\phantom{-\hat{A}_{1,1}}}_{n_1} & \underbrace{\phantom{\lambda\hat{B}_{1,2}-\hat{A}_{1,2}}}_{n_2} & & \underbrace{}_{n_{l+1}} \end{matrix}$$

where the matrices $\hat{A}_{i,i}$ are upper triangular matrices of full row rank, and the matrices $\hat{B}_{i,i+1}$ have full column rank. Since $\lambda\hat{B} - \hat{A}$ has only infinite elementary divisors, the $\hat{A}_{i,i}$ are square invertible and so is $\hat{A}_\infty$. Let us now introduce

$$\hat{N}_\infty \doteq \hat{B}_\infty \hat{A}_\infty^{-1}; \tag{50}$$

then $\hat{N}_\infty$ has exactly the same block structure as $\hat{B}_\infty$ since $\hat{A}_\infty^{-1}$ is upper triangular. Thus, $\hat{N}_\infty$ is nilpotent and we then have that

$$(\lambda\hat{B}-\hat{A})^{-1} = Z^* \cdot (\lambda\hat{B}_\infty - \hat{A}_\infty)^{-1} \cdot Q^* = Z^* \hat{A}_\infty^{-1} \cdot (I + \hat{N}_\infty \lambda + \hat{N}_\infty^2 \lambda^2 + \cdots + \hat{N}_\infty^l \lambda^l) \cdot Q^*. \tag{51}$$

The computation of $\hat{A}_\infty^{-1}$ is rather simple since it is triangular and so is the construction of $\hat{N}_\infty$. By Theorem 2 we find that the index $l+1$ obtained from the Kronecker canonical decomposition (i.e., the index of nilpotency of $\hat{N}$) equals the number of "stairs" $\hat{A}_{i,i}$ in (49), and hence also the index of nilpotency of $\hat{N}_\infty$.

The combination of (44), (47), and (51) now gives us

$$V_i = Z_{\text{left}} \hat{N}_\infty^i Q_{\text{right}} \qquad (i = 0, \cdots, l) \tag{52}$$

where

$$Z_{\text{left}} \doteq -[0, \cdots, 0, I_n] Z^* \hat{A}_\infty^{-1} \tag{53}$$

and

$$Q_{\text{right}} \doteq Q^* [0, \cdots, 0, I_n]^T. \tag{54}$$

Here Z_{left} and Q_{right} have dimensions $n \times dn$ and $dn \times n$, respectively.

Remark 4.1. If the unimodular matrix $U(\lambda)$ results from an embedding problem, then the construction of the previous section immediately yields a staircase form of the type (49). The possibility of choosing the diagonal blocks triangular in this embedding (see Remark 3.3) is thus appropriate here.

Remark 4.2. Since the index of nilpotency of $\hat{N}_\infty$ determines the number $l+1$ of coefficients V_i to be computed, trying to minimize l when dealing with the embedding problem is recommended. This is in fact done in the construction of Theorem 1: the lengths of the Jordan chains of the infinite elementary divisors—i.e., the number of stairs in the resulting staircase form (32)—is kept minimal, namely equal to the number of stairs in the staircase form (25) we are starting from. It is important to note here that not all V_i are necessarily nonzero, although the $\hat{N}^i$ and $\hat{N}_\infty^i$ matrices in (46) and (51) are nonzero for $i = 0, \cdots, l$. This thus means that l is in fact only an *upper bound* for the actual degree of $V(\lambda)$. This is, e.g., seen in the examples below.

We conclude this section with a summary of this procedure.

ALGORITHM INVERT.

(1) Construct the pencil $\lambda\hat{B} - \hat{A}$ defined by (42).
(2) Compute the staircase form of $\lambda\hat{B} - \hat{A}$ giving (49) with upper triangular $\hat{A}_{i,i}$ and compute $\hat{N}_\infty$ via (50).
(3) Compute Z_{left} and Q_{right} via (53) and (54).
(4) Compute the coefficients V_i of $V(\lambda)$ using (52).

5. Computational aspects. In the design of any numerical algorithm we are mainly concerned with two aspects: numerical reliability and computational speed.

As far as numerical precision is concerned, we can certainly say that the methods are based on the use of the staircase forms (25) or (49), which can be obtained by numerically stable algorithms [10], [11], [9], [12]. For the embedding problem this guarantees a rather good numerical behavior since the determination of $K_{c\infty}$, and subsequently of $Q(\lambda)$ via (35) and (24), does not seem to introduce any numerical difficulty. The method is, we believe, certainly to be preferred over the method using the Kronecker canonical form described in (21)–(22) or Eising's method [2], since these both require inverses of matrices that can be badly conditioned.

For the inversion problem the situation is somewhat different. There the use of the staircase form again avoids the use of the numerically sensitive Kronecker canonical form, but there is still an inversion problem involved. That this cannot be avoided is easily seen from the following recursions for the coefficients V_i of the inverse of a unimodular matrix:

$$V_0 = U_0^{-1}, \quad V_1 = -U_0^{-1} \cdot (U_1 V_0), \quad V_2 = -U_0^{-1} \cdot (U_2 V_0 + U_1 V_1), \cdots,$$
$$V_k = -U_0^{-1} \cdot \left(\sum_{i=0}^{k-1} U_{k-i} V_i \right). \tag{55}$$

If we *know* that the matrix $U(\lambda)$ is unimodular and that the degree of its inverse will be k, then this is probably the most direct (and also most reliable) method to compute the coefficients of $V(\lambda)$. But Algorithm Invert also provides a test for the unimodularity of $U(\lambda)$ and computes a (usually close) upper bound l for the degree k of its inverse. The algorithm is probably not much more sensitive than the mere application of (55), and it is certainly recommended for problems that are coming from an embedding since there $U(\lambda)$ is not directly available, whereas $\lambda \hat{B} - \hat{A}$ is.

Remark 5.1. It should be noted here that Eising also proposes a number of variants of his method which normally improve the numerical sensitivity of the problem, while allowing the embedding $U(\lambda)$ to have larger infinite elementary divisors than the minimum required. This is particularly important for the subsequent inversion problem where a trade-off between degree and sensitivity of the solution $V(\lambda)$ is pointed out by Eising [2].

As far as the computational complexity is concerned, we have already remarked that a cubic algorithm is available [1] for computing the staircase form of an arbitrary pencil, in contrast to the quartic methods that are available up to now [13], [10], [9], [6], [5]. For the embedding problem this decomposition constitutes the bulk of the work (namely $O(m^3d^3)$ flops) since the construction of $K \doteq K_{c\infty} \cdot V^*$ and $Q(\lambda)$ using (24) only require $O(m^2d^2n)$ flops and $O(md^2n(n-m))$ flops, respectively.

For the inversion problem we suppose first that it is connected to an embedding and, hence, that (49) is available. The computation of $\hat{N}_\infty$ and Z_{left} takes $O(n^3d^3)$ and $O(m^2d^2n)$ flops, respectively (Q_{right} is obtained at no cost). Starting with these data, the V_i are then computed recursively using

$$X_0 = Q_{\text{right}}, \quad V_0 = Z_{\text{left}} \cdot X_0, \quad \text{for } i = 1, \cdots, l\text{:} \quad X_i = \hat{N}_\infty \cdot X_{i-1}, \quad V_i = Z_{\text{left}} \cdot X_i \tag{56}$$

which takes $O(lm^2d^2n)$ flops for the total recursion. Here it is clear that it is very important to keep l as small as possible, since otherwise the complexity of this step may well become the larger part of the work (l may be as large as $md!$). If now the inversion problem is independent of an embedding, then the staircase form (49) has to be computed also which requires an additional $O(n^3d^3)$ flops. Moreover, one then has $m = n$.

We conclude this section with some numerical examples. The embedding problem largely relies on the staircase form, which has already been treated by various authors [13], [10], [1]. Therefore we restrict ourselves here to the inversion part.

Numerical examples. We give here some numerical examples of the Invert Algorithm. They were run on a VAX-750 computer with relative machine precision $EPS = 2^{-56} \approx 0.14 * 10^{-16}$. The notation is consistent with formulas (41)–(54). For brevity, we only list the nontrivial matrices. The computations were performed with the interactive matrix manipulation package MATLAB [7].

Example 1.

$$U(\lambda) \doteq \begin{bmatrix} 1 & \lambda & \lambda^2 \\ & 1 & \lambda \\ & & 1 \end{bmatrix}.$$

For $\lambda\hat{B}_\infty - \hat{A}_\infty = Q(\lambda\hat{B} - \hat{A})Z$ and $\hat{N}_\infty = \hat{B}_\infty\hat{A}_\infty^{-1}$ the following results were found up to 16 correct digits:

$$\hat{B}_\infty = \begin{bmatrix} 0 & 0 & -2\alpha & 0 & 0 & 0 \\ & & 0 & -2\alpha & -\alpha & 0 \\ & & & & \alpha & 0 \\ & & & & 0 & 1 \\ & & & & & 0 \\ & & & & & 0 \end{bmatrix}, \quad \hat{A}_\infty = \begin{bmatrix} -1 & & & & & \\ & -1 & & & & \\ & & -1 & & & \\ & & & +1 & & \\ & & & & -1 & \\ & & & & & -1 \end{bmatrix},$$

$$\hat{N}_\infty = \begin{bmatrix} 0 & 0 & 2\alpha & 0 & 0 & 0 \\ & & 0 & -2\alpha & \alpha & 0 \\ & & & & -\alpha & 0 \\ & & & & 0 & -1 \\ & & & & & 0 \\ & & & & & 0 \end{bmatrix},$$

$$Q = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ \alpha & 0 & 0 & 0 & \alpha & 0 \\ \alpha & 0 & 0 & 0 & -\alpha & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 0 & -\alpha & -\alpha & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\alpha & \alpha & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

where $\alpha = \sqrt{2}/2$ and $l = 2$. Straightforward computation of $U^{-1}(\lambda)$ using (52) gives

$$V(\lambda) = \begin{bmatrix} 1 & -\lambda & 0 \\ & 1 & -\lambda \\ & & 1 \end{bmatrix}.$$

Example 2.

$$U(\lambda) = \begin{bmatrix} 0 & \lambda^2 & 1 \\ 0 & 1 & 0 \\ 1 & \lambda + 7 & \lambda^2 + 7\lambda + 3 \end{bmatrix}.$$

In this case we obtained (up to 16 correct digits)

$$\hat{B}_\infty = \begin{bmatrix} 0 & -1 & 7 & 0 & 1 & 0 \\ & & -1 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 \\ & & & & 1 & 0 \\ & & & & & -1 \\ & & & & & 0 \end{bmatrix}, \quad \hat{A}_\infty = \begin{bmatrix} -1 & 0 & -3 & 0 & -7 & 0 \\ & 1 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 \\ & & & -1 & 0 & 0 \\ & & & & 1 & 0 \\ & & & & & -1 \end{bmatrix},$$

$$\hat{N}_\infty = \begin{bmatrix} 0 & -1 & 7 & 0 & 1 & 0 \\ & & -1 & 0 & 0 & 0 \\ & & & -1 & 0 & 0 \\ & & & & 1 & 0 \\ & & & & & 1 \\ & & & & & 0 \end{bmatrix},$$

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}.$$

Moreover, $l = 5$ and

$$V_0 = \begin{bmatrix} -3 & -7 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad V_1 = \begin{bmatrix} -7 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad V_2 = \begin{bmatrix} -1 & 3 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix},$$

$$V_3 = \begin{bmatrix} 0 & 7 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad V_4 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad V_5 = 10^{-17} * \begin{bmatrix} 0 & -0.3469 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Hence, when neglecting the term $\lambda^5 V_5$ (recall $EPS \approx 0.14 * 10^{-16}$) we indeed find the exact formula for the inverse of $U(\lambda)$, i.e.,

$$V(\lambda) = \begin{bmatrix} (-\lambda^2 - 7\lambda - 3) & (\lambda^4 + 7\lambda^3 + 3\lambda^2 - \lambda - 7) & 1 \\ 0 & 1 & 0 \\ 1 & -\lambda^2 & 0 \end{bmatrix}.$$

REFERENCES

[1] T. BEELEN, P. VAN DOOREN, AND M. VERHAEGEN, *A class of fast staircase algorithms for generalized state-space systems*, in Proc. American Control Conference, Seattle, WA, 1986, pp. 425–426.

[2] R. EISING, *Polynomial matrices and feedback*, IEEE Trans. Automat. Control, AC-30, (1985), pp. 1022–1025.

[3] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1959.

[4] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[5] V. KUBLANOVSKAYA, *AB-algorithm and its modifications for the spectral problems of linear pencils of matrices*, Numer. Math., 43 (1984), pp. 329–342.

[6] G. MIMINIS AND C. C. PAIGE, *An algorithm for pole assignment of time invariant multi-input linear system*, Proc. 21st IEEE Conference on Decision and Control, 1982, pp. 62–67.

[7] C. MOLER, *MATLAB user's guide*, Computer Science Department, University of New Mexico, Albuquerque, NM, 1980.

[8] N. MUNRO AND V. ZAKIAN, *Inversion of rational polynomial matrices*, Electronic Lett. 6 (1970), pp. 629–630.

[9] C. C. PAIGE, *Properties of numerical algorithms related to computing controllability*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 130–138.

[10] P. VAN DOOREN, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–141.

[11] ———, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 111–129.

[12] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

[13] ———, *Linear differential equations and Kronecker's canonical form*, in Recent Advances in Numerical Analysis, C. de Boor and G. Golub, eds., Academic Press, New York, 1978, pp. 231–265.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 1, January 1988

007

AN ANALOGUE OF THE SCHUR TRIANGULAR FACTORIZATION FOR COMPLEX ORTHOGONAL SIMILARITY AND CONSIMILARITY*

DIPA CHOUDHURY† AND ROGER A. HORN‡

Abstract. Any matrix $A \in M_n$ (the n-by-n complex matrices) can be triangularized by unitary similarity, i.e., there is a factorization $A = U\Delta U^*$, where $U \in M_n$ is unitary and $\Delta \in M_n$ is upper triangular; this is the well-known *Schur triangular factorization.* If $A\bar{A}$ has nonnegative spectrum, then A can also be triangularized by unitary consimilarity, i.e., there is a factorization $A = U\Delta U^T = U\Delta\bar{U}^{-1}$. We discuss the problem of triangularizing a given matrix by complex orthogonal similarity and consimilarity, i.e., factorizing $A = Q\Delta Q^T$ or $A = Q\Delta Q^*$, where $Q \in M_n$ is complex orthogonal.

Key words. triangularization, similarity, consimilarity, complex orthogonal matrices

AMS(MOS) subject classifications. 15A23, 15A21

Denote by $M_{m,n}$ the set of m-by-n complex matrices and set $M_n \equiv M_{n,n}$. We shall use P, Q to denote (complex) orthogonal matrices ($P, Q \in M_n$, $PP^T = QQ^T = I$). Denote the spectrum (set of eigenvalues) of a given $A = [a_{ij}] \in M_n$ by $\sigma(A)$, and define diag $(A) = \{a_{11}, a_{22}, \cdots, a_{nn}\}$, the set of main diagonal entries of A. The matrix diag $(x_1, x_2, \cdots, x_n) \in M_n$ denotes a diagonal matrix whose main diagonal entries are $x_1, x_2, \cdots, x_n$. A vector $x \in \mathbb{C}^n$ is called *isotropic* if $x^Tx = 0$, and *nonisotropic* if $x^Tx \neq 0$. A set of vectors $\{x_1, x_2, \cdots, x_n\} \in \mathbb{C}^n$ is *rectangular* if $x_i^Tx_j = 0$ for $i \neq j$; it is *rectanormal* if it is rectangular and $x_i^Tx_i = 1$ for all i. See [1], [5]–[7] for basic geometric and algebraic facts about rectangular and rectanormal sets, which are analogues for the symmetric bilinear form $b(x, y) = x^Ty$ of ordinary orthogonal and orthonormal sets with respect to the Hermitian form $h(x, y) = x^*y$.

A given matrix $A \in M_n$ is triangularized by similarity by a complex orthogonal matrix $Q \in M_n$ if $A = Q\Delta Q^T$, where $Q = [q_1\ q_2\ \cdots\ q_n]$ and

$$\Delta = \begin{bmatrix} \lambda_1 & & & * \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{bmatrix}$$

is upper triangular. Then $AQ = Q\Delta$ and hence $Aq_1 = \lambda_1 q_1$, i.e., q_1 is an eigenvector of A. Since q_1 is a column of Q, it is nonisotropic. Hence in order to have $A = Q\Delta Q^T$, the matrix A must have at least one nonisotropic eigenvector. But there are matrices, all of whose eigenvectors are isotropic.

Example. Let

$$A = \begin{bmatrix} 1 & -1 & 1 \\ 1/(2+\sqrt{2}) & 1 - i/\sqrt{2} & i/(2+\sqrt{2}) \\ -1/(2+\sqrt{2}) & -i/(2+\sqrt{2}) & 1 + i/\sqrt{2} \end{bmatrix}.$$

* Received by the editors September 15, 1987; accepted for publication October 1, 1987. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† Mathematical Sciences Department, Loyola College, Baltimore, Maryland 21210.

‡ Mathematical Sciences Department, The Johns Hopkins University, Baltimore, Maryland 21218.

The eigenvalues of this matrix are 1, $1 + i$, $1 - i$, and corresponding eigenvectors are

$$\begin{bmatrix} \sqrt{2}i \\ 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ i \end{bmatrix}, \quad \begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix};$$

they are determined up to a nonzero scalar factor and they are all isotropic. This matrix is diagonalizable, but it cannot be triangularized by a complex orthogonal matrix. See Corollary 3 for conditions that are sufficient to ensure that a diagonalizable matrix is orthogonally triangularizable.

THEOREM 1. *Let* $Z \equiv [z_1\ z_2\ \cdots\ z_n] \in M_n$ *be a nonsingular matrix that triangularizes a given matrix* $A \in M_n$, *i.e.,* $Z^{-1}AZ = \Delta_1$ *is an upper triangular matrix. Let* $Z_i = [z_1\ z_2\ \cdots\ z_i] \in M_{n,i}$ *for* $i = 1, \cdots, n$. *If* $\det (Z_i^T Z_i) \neq 0$ *for* $i = 1, 2, \cdots, n$, *then there exists a complex orthogonal matrix* $Q \in M_n$ *such that* $Q^T AQ = \Delta$ *is upper triangular.*

Proof. Because of the nonsingularity assumptions on each $Z_i^T Z_i$, there exists a rectanormal set $\{q_1, q_2, \cdots, q_n\} \subset \mathbb{C}^n$ that is triangularly equivalent to the set of columns of Z, i.e., there exists a nonsingular upper triangular matrix B such that

$$Z = QB \tag{2}$$

(see [1, Lemma 2.6]). By assumption, $Z^{-1}AZ = \Delta_1$ is upper triangular. Using (2) we have $\Delta_1 = Z^{-1}AZ = (QB)^{-1}A(QB) = B^{-1}(Q^T AQ)B$, i.e., $Q^T AQ = B\Delta_1 B^{-1} \equiv \Delta$ is upper triangular. □

There are many possible triangularizations of a given matrix A as $A = Z\Delta Z^{-1}$; some similarity matrices Z may satisfy the condition $\det (Z_i^T Z_i) \neq 0$ for $i = 1, \cdots, n$ and some may not. Thus, some unitary triangularizations of a given matrix A may lead to an orthogonal triangularization by the process described in the theorem and some may not.

Example. Let

$$A = \begin{bmatrix} 1 & 1+i \\ i-1 & -1 \end{bmatrix}.$$

The unitary matrix

$$U = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$$

triangularizes A and

$$U^* AU = \begin{bmatrix} i & 2i \\ 0 & -i \end{bmatrix}.$$

Both columns of U are isotropic, so, in particular, it fails to satisfy $\det (U_1^T U_1) \neq 0$.

Now consider the real orthogonal (and hence complex orthogonal and unitary) matrix

$$Q = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix},$$

which also triangularizes A:

$$Q^*AQ=\begin{bmatrix}-i & 2\\ 0 & i\end{bmatrix}.$$

Since Q is a real nonsingular matrix it satisfies $\det(Q_i^TQ_i)\neq 0$ for $i=1, 2$. Thus, the given matrix A can be upper triangularized by a complex orthogonal matrix Q.

COROLLARY 3. *Let $x_1, x_2, \cdots, x_n \in \mathbb{C}^n$ be eigenvectors of a given matrix $A\in M_n$ and let $X_i=[x_1\ x_2\ \cdots\ x_i]\in M_{n,i}$ for $i=1, \cdots, n$. If $\det(X_i^TX_i)\neq 0$ for $i=1, \cdots, n$, then there exists a complex orthogonal matrix $Q\in M_n$ such that $Q^TAQ=\Delta$ is upper triangular.*

Proof. Let $X\equiv[x_1\ x_2\ \cdots\ x_n]\in M_n$, which is nonsingular since $\det(X_n^TX_n)\neq 0$. Thus, the matrix X diagonalizes A; in particular, it triangularizes A. By Theorem 1 there exists a complex orthogonal matrix $Q\in M_n$ such that Q^TAQ is upper triangular. □

LEMMA 4. *Let $A\in M_n$ be such that*

(a) *$A\bar{A}$ is real,*

(b) *$A+\bar{A}$ has only real eigenvalues, and*

(c) *$A-\bar{A}$ has only imaginary eigenvalues.*

Then there exists a complex orthogonal matrix $Q\in M_n$ such that Q^TAQ is upper triangular.

Proof. Let $A=C+iD$, where C and D are real. Since $A\bar{A}$ is real, $A\bar{A}=\overline{A\bar{A}}=\bar{A}A$, i.e., $(C+iD)(C-iD)=(C-iD)(C+iD)$, from which it follows that $CD=DC$. By assumption, $A+\bar{A}=C+iD+C-iD=2C$ has only real eigenvalues and $A-\bar{A}=C+iD-C+iD=2iD$ has only imaginary eigenvalues, i.e., D has only real eigenvalues.

Since C and D commute and have only real eigenvalues, they can be simultaneously triangularized by a real unitary matrix Q, which is therefore an orthogonal matrix [3, Thm. (2.3.6)]. Thus, $A=C+iD=Q^T\Delta_1Q+iQ^T\Delta_2Q=Q^T(\Delta_1+i\Delta_2)Q=Q^T\Delta Q$, where Δ_1, Δ_2, and $\Delta\equiv\Delta_1+i\Delta_2$ are all upper triangular. □

Although $A\bar{A}$ need not be real in general, it is always similar to a real matrix (in fact, to the square of a real matrix). The following example illustrates the hypotheses of the lemma.

Example. Let

$$A=\begin{bmatrix}i-1 & 2\\ -1 & 2+i\end{bmatrix}.$$

Then

$$A\bar{A}=\begin{bmatrix}0 & 2\\ -1 & 3\end{bmatrix}$$

is real. The eigenvalues of $A+\bar{A}$ are 0, 2, and the eigenvalues of $A-\bar{A}=2iI$ are all imaginary. Thus, this matrix is orthogonally triangularizable by Lemma 4. Since A has real independent eigenvectors $[2\ \ 1]^T$ and $[1\ \ 1]^T$, it also satisfies the hypotheses of Corollary 3.

The hypotheses of the preceding lemma are sufficient but not necessary for orthogonal triangularization. Consider the following example.

Example. Let

$$A=\begin{bmatrix}2+i & -2(1+i)\\ 2+i & 0\end{bmatrix}.$$

Then

$$A\bar{A} = \begin{bmatrix} -(2i+1) & -6+2i \\ 5 & -6+2i \end{bmatrix},$$

which is not real. The eigenvalues of $A + \bar{A}$ are $2 \pm 2\sqrt{3}i$, which are not real, and the eigenvalues of $A - \bar{A}$ are $i \pm \sqrt{7}$, which are not pure imaginary. Nevertheless, A is orthogonally triangularizable by Corollary 3 because it has distinct eigenvalues and both eigenvectors are nonisotropic.

We have so far found some conditions that are sufficient for a given matrix $A \in M_n$ to be triangularizable by orthogonal similarity. We have not yet found useful conditions that are both necessary and sufficient for the triangularizability of a given matrix by orthogonal similarity. The difficulty in the study of orthogonal similarity arises from dealing with isotropic eigenvectors. But if we consider orthogonal consimilarity, i.e.,

$$A \to QA\overline{Q^{-1}} = QAQ^*,$$

where Q is a complex orthogonal matrix, there are some useful ways to deal with isotropic coneigenvectors. We now discuss triangularization of a matrix by complex orthogonal consimilarity.

DEFINITION 5. A matrix $A \in M_n$ is said to be *contriangularizable* if there exists a nonsingular $R \in M_n$ such that $R^{-1}A\bar{R}$ is upper triangular. A nonzero vector $x \in \mathbb{C}^n$ such that $A\bar{x} = \lambda x$ is said to be a *coneigenvector* of A; the scalar λ is a *coneigenvalue* of A.

Every matrix has at least one eigenvalue and has only finitely many distinct eigenvalues, but the situation is fundamentally different for coneigenvalues and coneigenvectors. If $A \in M_n$ and $A\bar{x} = \lambda x$, then

$$e^{-i\theta}A\bar{x} = A(\overline{e^{i\theta}x}) = e^{-i\theta}\lambda x = (e^{-2i\theta}\lambda)(e^{i\theta}x)$$

for all $\theta \in \mathbb{R}$. Thus, if λ is a coneigenvalue of A, then so is $e^{-2i\theta}\lambda$ for all $\theta \in \mathbb{R}$. On the other hand, if $A\bar{x} = \lambda x$ then

$$A\bar{A}x = A(\overline{A\bar{x}}) = A(\overline{\lambda x}) = \bar{\lambda}(A\bar{x}) = \bar{\lambda}\lambda x = |\lambda|^2 x,$$

so a complex scalar λ is a coneigenvalue of A only if the nonnegative real number $|\lambda|^2$ is an eigenvalue of $A\bar{A}$. In fact, A has a coneigenvector if and only if some eigenvalue of $A\bar{A}$ is nonnegative. For example,

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \qquad A\bar{A} = -2I$$

has no nonnegative eigenvalues, and hence the matrix A has no coneigenvalues.

Thus, a matrix may have infinitely many distinct coneigenvalues or it may have no coneigenvalues at all. For more information on consimilarity, coneigenvalues and contriangularization see [2].

DEFINITION 6. The *field of values* of $A \in M_n$ is $F(A) = \{x^*Ax: x^*x = 1, x \in \mathbb{C}^n\}$. For basic properties of the field of values see Chapter 1 of [4].

LEMMA 7. *Let $A \in M_n$ be given. If $0 \notin F(A)$ then $0 \notin F(C^*AC)$ for every nonsingular $C \in M_n$.*

Proof. The assumption that $0 \notin F(A)$ means that $y^*Ay \neq 0$ for all y such that $y^*y = 1$. If $0 \in F(C^*AC)$, then there exists $x \in \mathbb{C}^n$ such that $x^*x = 1$ and $x^*(C^*AC)x = 0$. Since C is nonsingular, $Cx \neq 0$ and we may set $z = Cx/((Cx)^*(Cx))^{1/2}$. Then $z^*z = 1$ and we have $z^*Az = 0$, which contradicts our hypothesis that $0 \notin F(A)$. □

Since the field of values $F(A)$ is a convex set that contains all the eigenvalues of A, assuming that $0 \notin F(A)$ is stronger than assuming that A is nonsingular. In particular, $0 \notin F(A)$ implies that 0 is not in the convex hull of the spectrum of A.

LEMMA 8. *Let $A \in M_n$ be a given matrix such that $0 \notin F(A)$. If $A\bar{A}$ has a nonnegative eigenvalue, then A has a nonisotropic coneigenvector.*

Proof. Since $A\bar{A}$ has a nonnegative eigenvalue, A has a coneigenvector x, which we may take to be a unit vector. Thus, $A\bar{x} = \lambda x$, and consequently $x^T A\bar{x} = \lambda x^T x$. If x is isotropic then $x^T A\bar{x} = (\bar{x})^* A(\bar{x}) = 0$, which contradicts the assumption that $0 \notin F(A)$. Hence x is nonisotropic. □

DEFINITION 9. A matrix $A \in M_n$ has the *orthogonally inheritable nonisotropic coneigenvector property* if whenever $Q \in M_n$ is orthogonal and

$$QAQ^* = \left[\begin{array}{cc|c} * & * & \\ & \ddots & * \\ 0 & * & \\ \hline 0 & & \tilde{A} \end{array}\right], \qquad \tilde{A} \in M_k, \quad 1 \leqq k \leqq n$$

is a partial orthogonal contriangularization of A, then some coneigenvector of $\tilde{A}$ is nonisotropic.

THEOREM 10. *If $A \in M_n$ has the orthogonally inheritable nonisotropic coneigenvector property, there exists a complex orthogonal matrix Q such that QAQ^* is upper triangular.*

Proof. Apply the same step-by-step reduction used to prove Schur's triangularization theorem (see [3, Thm. 2.3.1]), but consider coneigenvectors instead of eigenvectors and construct an orthogonal matrix instead of a unitary matrix. □

COROLLARY 11. *Let $A \in M_n$ be such that $0 \notin F(A)$. There exists a complex orthogonal matrix $Q \in M_n$ such that QAQ^* is upper triangular if and only if all the eigenvalues of $A\bar{A}$ are nonnegative.*

Proof. If there is nonsingular $R \in M_n$ such that $R^{-1}A\bar{R} = \Delta$ is upper triangular, then the main diagonal entries of $\Delta\bar{\Delta} = (R^{-1}A\bar{R})(R^{-1}A\bar{R}) = R^{-1}(A\bar{A})R$ are nonnegative and are the eigenvalues of $A\bar{A}$. Conversely, we shall show that if $0 \notin F(A)$ and $\sigma(A\bar{A}) \geqq 0$, then A has the orthogonally inheritable coneigenvector property. Let $Q \in M_n$ be orthogonal and suppose that

$$QAQ^* = \left[\begin{array}{cc|c} * & * & \\ & \ddots & * \\ 0 & * & \\ \hline 0 & & \tilde{A} \end{array}\right], \qquad \tilde{A} \in M_k, \quad 1 \leqq k \leqq n.$$

Then

$$QA\bar{A}Q^T = QAQ^*\overline{QA}Q^T = (QAQ^*)(\overline{QAQ^*}) = \left[\begin{array}{cc|c} * & * & \\ & \ddots & * \\ 0 & * & \\ \hline 0 & & \tilde{A}\bar{\tilde{A}} \end{array}\right].$$

Since the eigenvalues of $A\bar{A}$ are nonnegative, all the eigenvalues of $\tilde{A}\bar{\tilde{A}}$ are also nonnegative. Lemma 7 guarantees that $0 \notin F(QAQ^*)$, and hence $0 \notin F(\tilde{A})$. Thus, $\tilde{A}$ has a nonisotropic coneigenvector by Lemma 8, i.e., A has the orthogonally inheritable nonisotropic coneigenvector property. The conclusion follows from Theorem 10. □

It is known [2] that if $A \in M_n$, then $\sigma(A\bar{A}) \geqq 0$ if and only if there is a unitary $U \in M_n$ such that UAU^T is upper triangular. Thus, the preceding corollary says that if $0 \notin F(A)$, then $A \in M_n$ is unitarily contriangularizable if and only if it is orthogonally contriangularizable.

DEFINITION 12. A matrix $A \in M_n$ is said to be *condiagonalizable* if there exists a nonsingular $R \in M_n$ such that $R^{-1}A\bar{R}$ is diagonal.

If $A \in M_n$ is condiagonalizable and $R^{-1}A\bar{R} = \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$, then $A\bar{R} = R\Lambda$. If $R = [r_1\ r_2\ \cdots\ r_n]$ with each $r_i \in \mathbb{C}^n$, this identity says that $A\bar{r}_i = \lambda_i r_i$ for $i = 1, \cdots, n$. The identity $A\bar{R} = R\Lambda$ (without any assumption about the nonsingularity of R) says that every nonzero column of the matrix R is a coneigenvector of A. Since the columns of R are independent if and only if R is nonsingular, a matrix $A \in M_n$ is condiagonalizable if and only if it has n independent coneigenvectors [2].

The following corollary shows that any positive definite matrix can be condiagonalized by a complex orthogonal matrix.

COROLLARY 13. *Let $A \in M_n$ be given. Then A is positive definite if and only if there exists a complex orthogonal $Q \in M_n$ such that $QAQ^* = \Lambda$ is diagonal with positive main diagonal elements, i.e., $A = P\Lambda P^*$, where $P \equiv Q^T$ is complex orthogonal.*

Proof. If $A \in M_n$ is positive definite, then $A\bar{A}$ is similar to $A^{-1/2}A\bar{A}A^{1/2} = A^{1/2}\bar{A}A^{1/2}$. Since $\bar{A}$ is positive definite, $A^{1/2}\bar{A}A^{1/2}$ (and hence also $A\bar{A}$) has positive eigenvalues. Since $F(A)$ is the convex hull of $\sigma(A)$ and $\sigma(A) > 0$, $0 \notin F(A)$. By Corollary 11 there exists a complex orthogonal matrix Q such that $QAQ^* = \Delta$ is upper triangular. Since QAQ^* is Hermitian, Δ must be diagonal, i.e., $QAQ^* \equiv \Lambda$ is diagonal and positive definite, so its diagonal entries must be positive. The converse assertion follows immediately. □

REFERENCES

[1] D. CHOUDHURY AND R. A. HORN, *An analog of the Gram–Schmidt algorithm for complex bilinear forms and diagonalization of complex symmetric matrices*, Technical Report No. 454, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD, January 2, 1986.

[2] Y. P. HONG AND R. A. HORN, *On the reduction of a matrix to triangular or diagonal form by consimilarity*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 80–86.

[3] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[4] ———, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1988.

[5] I. KAPLANSKY, *Linear Algebra and Geometry*, Chelsea, New York, 1974.

[6] O. T. O'MEARA, *Introduction to Quadratic Forms*, third corrected printing, Springer-Verlag, Berlin, 1973.

[7] E. SNAPPER AND R. TROYER, *Metric Affine Geometry*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 1, January 1988

008

PARALLEL PROCESSING IN THE ADAPTIVE CONTROL OF LINEAR SYSTEMS*

ROBERTO CRISTI†

Abstract. The implementation of a direct adaptive control algorithm using parallel processing techniques is discussed. The controller presented is hybrid in nature (continuous time feedback and discrete time gain adjustment) and the recursive least squares identification is implemented using a well-known algorithm based on the Givens rotation.

Key words. parallel processing, adaptive control, systolic arrays

AMS(MOS) subject classification. 93

1. Introduction. The problem of adaptively controlling plants with uncertainties has received particular attention during the last 15 years. This increase in popularity is caused by the desire to obtain the best possible performances in spite of uncertainties of the system and/or changing operating conditions, and the extraordinary increase in computer technology witnessed during the last two decades.

At the present time, the long-standing questions of global and local stability have been answered [1], [2], while we are close to a better understanding in regard to problems related to robustness in the presence of output disturbances [3] and unmodeled dynamics [4]. Also, the limitations of early results to minimum phase systems have recently been overcome in [5], [6], substantiating an early conjecture [7] that global stability can be guaranteed provided all modes of the plants are excited.

All these refinements of adaptive control algorithms have been obtained at the expense of added complexity with respect to early schemes. For example, it is well recognized that algorithms based on Recursive Least Squares converge much faster than simpler Projection Algorithms [8]. The result is an improved convergence rate of the estimated parameters, and consequently better tracking performances for systems with time varying parameters, and better robustness [9] at the expense of the need of more computing power.

Also, a class of adaptive controllers for nonminimum phase systems presented in [6] requires the estimate of a redundant number of parameters.

From this list of recent results, the need for adequate computing capabilities is evident. Systems with relatively small time constants need fast sampling rates, and complex adaptive techniques might require computing speeds inadequate to available microcomputers.

A possible answer to this problem can be found in parallel computing structures, such as systolic arrays or wavefront arrays. In this report the adaptive algorithm presented in [6] based on recursive least squares with periodic covariance resetting is redesigned, in order to make it suitable to implementation on a VLSI chip. The motivation is to be able to obtain a throughput of the data acceptable for high speed operations of systems with a large number of parameters to be estimated.

The basic idea is common to least squares algorithms found in signal processing literature [10], [11], where the desired parameters are estimated from an upper triangular

* Received by the editors June 12, 1987; accepted for publication October 1, 1987. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† Electrical and Computer Engineering Department, Naval Postgraduate School, Monterey, California 93943. This work has been supported by the NPS Foundation for Research, contract RYEHK, 1987.

factorization (called QR [18]) of the data matrix. The difference in adaptive control is that the estimator has to operate recursively on subsequent blocks of data, so that the estimated parameters converge asymptotically to the respective correct values. Asymptotic convergence is guaranteed by proper initialization at each block and persistency of excitation of the external input.

This report is divided as follows: the Model Reference Adaptive Control problem is recalled in § 2, with its hybrid implementation in § 3. Parameter estimation with its parallel implementation are the subjects of §§ 4 and 5, while global stability and performances considerations with regard to the block processing and hybrid approaches are given in §§ 6 and 7.

2. Adaptive control of linear systems. The dynamics of a linear Single Input Single Output (SISO) system can be modeled by the differential equation

$$p(D)y(t) = r(D)u(t) \tag{2.1}$$

where y, $u{:}R^+ \to R$ represent output and input signals, respectively, and p and r are polynomials in the differential operator $D = d/dt$ as

$$\begin{aligned} p(D) &= D^n + a_1^{n-1} + \cdots + a_n, \\ r(D) &= K_p(D^{n-m} + \cdots + r_n). \end{aligned} \tag{2.2}$$

The following assumptions on the plant (2.1) will be made throughout.

(A1) The order of the plant n and its relative degree m are known to the designer.

(A2) The values of the plant coefficients r_i, p_j, K_p are unknown while the sign of the leading coefficient K_p and a lower bound on its magnitude are known to the designer.

(A3) The polynomial $r(D)$ is Hurwitz (i.e., plant minimum phase).

(A4) $r(D)$ and $p(D)$ are mutually coprime polynomials.

With the above assumptions the aim of the adaptive controller is to determine a control input $u(t)$ so that the output of the plant $y(t)$ tracks the output of a linear model $y_m(t)$ defined as

$$p^*(D)y_m(t) = v(t). \tag{2.3}$$

The reference model (2.3) with transfer function $1/p^*(s)$ represents the desired asymptotic performance and p^* is assumed to be an arbitrary Hurwitz polynomial of degree m (the relative degree of the plant).

Remark. In view of recent results [5], [6], the assumptions (A1)–(A3) above can be considerably relaxed in the sense that the plant can be assumed to be just of order n (known) and strictly proper. However, this gain in generality is obtained by more complex adaptive controllers that are beyond the scope of this report.

The structure of the adaptive controller can be easily determined on the basis of the fixed control problem, by which we seek arbitrary pole placement with the control input $u(t)$ defined by the differential equation

$$q(D)u(t) = h(D)y(t) + k(D)u(t) + q(D)v(t), \tag{2.4}$$

or alternatively as

$$u(t) = h(D)\tilde{y}(t) + k(D)\tilde{u}(t) + v(t). \tag{2.5}$$

The polynomials q, h, k above have degrees n, $n-1$, $n-1$, respectively. In particular $q(D)$ is an arbitrary nth degree Hurwitz polynomial (the observer polynomial) and in the fixed control case (i.e., plant dynamics assumed to be fully known) $h(D)$ and $k(D)$ are determined by the Diophantine Equation

$$h(D)r(D)+k(D)p(D)=q(D)[p(D)-K_p^{-1}r(D)p^*(D)]. \tag{2.6}$$

The existence of a unique solution h, k of (2.6) is guaranteed by the assumed coprimeness of $r(D)$ and $p(D)$ (assumption (A4) above) and by choosing h and k both of degree $n-1$. In the expression (2.5) $\tilde{u}(t)$, $\tilde{y}(t)$ represent filtered input and output of the plant defined by the differential equations

$$q(D)\tilde{y}(t)=y(t), \qquad q(D)\tilde{u}(t)=u(t). \tag{2.7}$$

3. Adaptive controller and hybrid structure. When the plant dynamics are not known, the compensator parameters (i.e., the coefficients of the polynomials $h(D)$ and $k(D)$) have to be estimated recursively from input and output measurements of the plant. The particular structure we consider is hybrid in nature, in the sense that the compensator parameters are updated on a discrete time basis, from samples taken on the signals of the loop. For this we define the sampling time sequence

$$\{t_k^i; k=0,1,\cdots;i=0,1,\cdots,N-1\}$$

as

$$t_k^i=kNT+iT+\tau_k. \tag{3.1}$$

The reasons beyond this definition concern global stability problems and are fully discussed in [6] and [12]. In particular the adaptive gains are updated at times $\{t_k^0\}$ on the basis of the samples taken at times $\{t_{k-1}^i;\ i=0,\cdots,N-1\}$ called the kth time block.

The sequence τ_k is a random independently and identically distributed sequence uniformly distributed on any arbitrarily small interval. It is included in (3.1) in order to guarantee (with probability one) observability of continuous time modes from sampled values of the loop signals [12].

The compensator parameters are estimated directly from the loop signals, based on manipulation of the Diophantine Equation and definition of the partial state $z(t)$ [13] by which we can write (2.1) as

$$p(D)z(t)=u(t), \qquad y(t)=r(D)z(t). \tag{3.2}$$

The input and output signals u, y can be viewed as linear combinations of derivatives of the partial state $z(t)$. It is easy to see that $z(t)$ with its first $n-1$ derivatives constitute the entries of the state $\underline{x}(t)$ of the controllable canonical form realization of the plant (2.1) [13]. By operating left- and right-hand sides of (2.6) on $z(t)$, and keeping (3.2) in mind, we can relate the polynomials $h(D)$ and $k(D)$ to the signals $u(t)$ and $y(t)$ as

$$h(D)y(t)+k(D)u(t)=q(D)u(t)-q(D)K_p^{-1}p^*(D)y(t). \tag{3.3}$$

In order to have signals obtainable with proper (and therefore physically realizable) transformations, define $\bar{y}$, $\bar{u}$, u_f as

$$\begin{aligned} &q(D)p^*(D)\bar{y}(t)=y(t),\\ &q(D)p^*(D)\bar{u}(t)=u(t),\\ &p^*(D)u_f(t)=u(t) \end{aligned} \tag{3.4}$$

and write (3.3) as

$$h(D)\bar{y}(t)+k(D)\bar{u}(t)+K_p^{-1}y(t)=u_f(t) \tag{3.5}$$

or in compact form

$$\begin{aligned}&\underline{\theta}^{*T}\underline{\phi}(t)=u_f(t),\\ &\underline{\theta}=[h_1,\cdots,h_{n-1},k_1,\cdots,k_{n-1},K_p^{-1}],\\ &\underline{\phi}(t)=[\bar{y}(t),\cdots,D^{n-1}\bar{y}(t),\bar{u}(t),\cdots,D^{n-1}\bar{u}(t),y(t)]\end{aligned} \tag{3.6}$$

with h_i, k_j being the coefficients of the respective polynomials. At the sampling instants t_k^i (3.6) can be treated as a relation between sequences $u_f(t_k^i)$ and $\underline{\phi}(t_k^i)$, and the parameter vector $\underline{\theta}^*$ can be recursively estimated on a discrete time basis.

4. Parameter estimation. Several techniques in the literature [8], [14] allow estimates of $\underline{\theta}^*$ to be computed on line. The particular one we consider in this report is based on the Recursive Least Squares (RLS) with Covariance Resetting introduced by several authors [6], [15]. In particular, by this algorithm the sequence of plant estimates $\hat{\underline{\theta}}_k$ at times t_k^0 (beginning of the kth time block) is computed in the form

$$\hat{\underline{\theta}}_{k+1}=F(\hat{\underline{\theta}}_k/\underline{\phi}(t_k^i),u_f(t_k^i),\text{ for } i=0,\cdots,N-1) \tag{4.1}$$

for some function F given below, which depends on the previous parameter estimate and the data in the block.

A recursive version of this estimation is given by the well-known recursions

$$\begin{aligned}&\hat{\underline{\theta}}_k^i=\hat{\underline{\theta}}_{k-1}^{i-1}-\frac{P_k^{i-2}\underline{\phi}(t_k^{i-1})e(t_k^{i-1})}{1+\underline{\phi}(t_k^{i-1})^TP_k^{i-2}\underline{\phi}(t_k^{i-1})},\\ &P_k^i=P_k^{i-1}-\frac{P_k^{i-1}\underline{\phi}(t_k^i)\underline{\phi}(t_k^i)^TP_k^{i-1}}{1+\underline{\phi}(t_k^i)^TP_k^{i-1}\underline{\phi}(t_k^i)},\\ &\underline{\theta}_k^0=\hat{\underline{\theta}}_{k-1}^N=\hat{\underline{\theta}}_k,\\ &P_k^{-1}=\sigma_0^2I,\\ &e(t_k^i)=u_f(t_k^i)-\hat{\underline{\theta}}_k^i\underline{\phi}(t_k^i)\end{aligned} \tag{4.2}$$

where σ_0 is an arbitrary positive constant. In (4.2) the quantities $\hat{\underline{\theta}}_k^i$, P_k^i are computed at each instant t_k^i defined in (3.1).

The relevant feature of the RLS algorithm with Covariance Resetting is the fact that the covariance matrix P_k^i is periodically reset to its initial condition. This prevents P_k^i from decaying to zero, as in the standard least squares algorithm, which would mean a loss of sensitivity as time increases.

For the parallel implementation using systolic arrays we compute $\hat{\underline{\theta}}_k$ directly from the minimization of a proper quadratic cost function, disregarding the estimates $\hat{\underline{\theta}}_k^i$, $i\neq 0, N-1$ within the time block. In particular the cost function

$$V_k(\underline{\theta})=\sum_{j=0}^{N-1}|u_f(t_k^i)-\underline{\theta}^T\underline{\phi}(t_k^i)|^2+\|\underline{\theta}-\hat{\underline{\theta}}_k\|^2\sigma_0^{-2} \tag{4.3}$$

on which RLS estimation is based [8] is used to define implicitly the estimate $\hat{\underline{\theta}}_{k+1}$ at instants t_{k+1}^0 as

$$V_k(\hat{\underline{\theta}}_{k+1})=\min_\theta V_k(\underline{\theta}). \tag{4.4}$$

From the definition of V_k in (4.3) we can see that the parameter σ_0 weights the confidence we have on the initial (for each block) estimate $\underline{\hat{\theta}}_k$.

5. Recursive estimation by systolic arrays. In this section we discuss the recursive computation of the sequence of parameter estimates $\underline{\hat{\theta}}_k$ from (4.3) using systolic arrays techniques.

Simple algebraic manipulations on (4.3) and definition of the Euclidean norm yield $\underline{\hat{\theta}}_{k+1}$ as the solution of the linear algebraic equation

$$\begin{bmatrix} \underline{\bar{\phi}}(t_k^{N-1})^T \\ \vdots \\ \underline{\bar{\phi}}(t_k^0)^T \\ I \end{bmatrix} \underline{\theta} = \begin{bmatrix} \bar{u}_f(\underline{t}k^{N-1}) \\ \vdots \\ \bar{u}_f(\underline{t}k^0) \\ \underline{\hat{\theta}}_k \end{bmatrix} \tag{5.1}$$

where the equality is in the least squares sense (minimum square error). The signals in (5.1) are normalized by the constant σ_0 as

$$\underline{\bar{\phi}}(t) = \underline{\phi}(t)/\sigma_0, \qquad \bar{u}_f(t) = u_f(t)/\sigma_0. \tag{5.2}$$

The solution to (5.1) always exists and is unique since the leftmost matrix has full rank due to the identity block I.

A common way to obtain the least squares solution of (5.1) is by a QR factorization of the leftmost matrix, which can be carried out using parallel processing and systolic arrays techniques. The basic idea, as presented originally in [11], is to transform a "tall" matrix into an upper triangular one by successive linear combinations of pairs of rows and force zeros into desired positions. The transformation can be considered as a succession of elementary vector rotations (the Givens rotation [16]). Details of the algorithm can be found in numerous references [10], [11].

For any $N \times N$ matrix A the Givens rotation is characterized by a square matrix

$$Q(p, q) = \operatorname{diag}(I_1, \gamma(p, q), I_2)$$

associated to a pair of indices $(p, q) \in [2, N] \times [2, N]$ with I_1, I_2 the identity matrices of sizes $q - 2$ and $N - q$, respectively, and $\gamma(p, q)$ a 2×2 matrix defined as

$$\gamma(p, q) = \begin{bmatrix} c(p, q) & s(p, q) \\ -s(p, q) & c(p, q) \end{bmatrix} \tag{5.3}$$

with

$$c(p, q) = \frac{a_{q-1,p}}{a_{q-1,p}^2 + a_{q,p}^2}, \qquad s(p, q) = \frac{a_{p,q}}{a_{q-1,p}^2 + a_{q,p}^2}, \tag{5.4}$$

$a_{i,j}$ being the entries of the given matrix A. It is easy to see that left multiplication of A by Q forces a one and a zero to appear as follows:

$$Q(p, q)A = \begin{matrix} \\ \\ q \\ \\ \end{matrix} \begin{matrix} \begin{matrix} \quad p \end{matrix} \\ \begin{bmatrix} x & x & x \\ x & 1 & x \\ x & 0 & x \\ x & x & x \end{bmatrix} \end{matrix} \tag{5.5}$$

where x indicates other elements of the matrix.

As an example of application, and to make the operations described more transparent, we can see that an almost upper triangular matrix can be made upper triangular by a succession of Givens rotations as follows:

$$\begin{bmatrix} r'_{11} & r'_{12} & r'_{13} \\ 0 & r'_{22} & r'_{23} \\ 0 & 0 & r'_{33} \\ 0 & 0 & 0 \end{bmatrix} = Q(3,4)Q(2,3)Q(1,2) \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 \\ r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix}. \tag{5.6}$$

This can be applied to the solution of (5.1) within the kth time block to obtain the following algorithm:

initialize at t_k^0: $R_k(-1) = I$; $\beta_k(-1) = \hat{\underline{\theta}}_k$;
compute at each t_k^i, $i = 0, \cdots, N-1$

$$\begin{aligned} R_k(i/i-1) &= \begin{bmatrix} \underline{\phi}(t_k^i)^T \\ R_k(i-1) \end{bmatrix}, \\ \beta_k(i/i-1) &= \begin{bmatrix} u_f(t_k^i) \\ \beta_k(i-1) \end{bmatrix}, \\ \begin{bmatrix} R_k(i) \\ 0 \end{bmatrix} &= Q_k(i) R_k(i/i-1), \\ \begin{bmatrix} \beta_k(i) \\ x \end{bmatrix} &= Q_k(i) \beta_k(i/i-1), \end{aligned} \tag{5.7}$$

output $R_k(N-1)$ at each t_k^{N-1}.

In the above algorithm the matrices $R_k(i)$, $i = 0, \cdots, N-1$ are square upper triangular of dimensions $2n \times 2n$ (since the number of parameters in the adaptive algorithm is $2n$), $R_k(i/i-1)$ are $(2n+1) \times 2n$ and similarly $\beta_k(i)$ and $\beta_k(i/i-1)$ are $2n \times 1$ and $(2n+1) \times 1$, respectively. The matrix $Q_k(i)$ indicates the rotation matrix at each t_k^i, as the product of the Q matrices in the example (5.6).

After the transformation of the data matrix into the upper triangular $R_k(N-1)$ in the above algorithm, the parameter estimate $\hat{\underline{\theta}}_{k+1}$ can be computed from the system of equations

$$R_k(N-1)\hat{\underline{\theta}}_{k+1} = \beta_k(N-1) \tag{5.8}$$

solvable by successive substitutions due to the triangular nature of $R_k(N-1)$.

The two operations (5.7) and (5.8) can be carried out by two distinct systolic array processors, as shown in Fig. 1. Processor P1 performs the triangularization operations (5.7) within the time block $\{t_k^i, i = 0, \cdots, N-1\}$, while P2 carries the linear solution as in (5.8) and it operates at the end of the time block.

The structure of the processor P1 is shown in Fig. 2, and the cell operations are defined in Table 1. In the actual implementation particular care has to be devoted to the correct timing of the data, due to the space-time nature of the structure. As shown in Fig. 2 the regression vector is input to the array in a skewed fashion, by properly delaying its entries. The reset command must provide for data output at the end of the block, and array initialization for the next block computation.

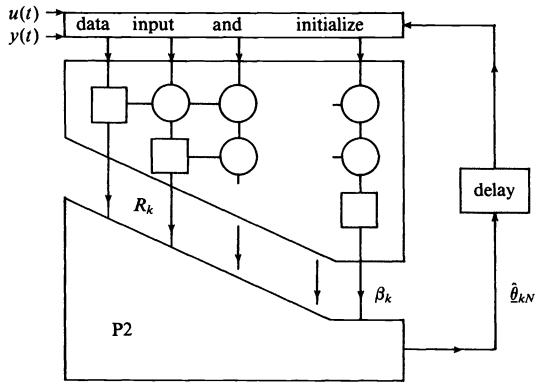


FIG. 1. *Two processor structure.*

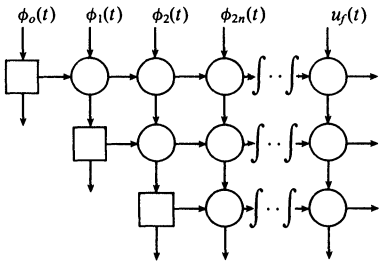


FIG. 2. *Systolic array for triangularization.*

6. Global stability and tracking capabilities. The control input to the plant is determined as

$$u(t) = \hat{h}_k(D)\tilde{y}(t) + \hat{k}_k(D)\tilde{u}(t) + v(t) \tag{6.1}$$

for $t \in [t_k^0, t_{k+1}^0)R$, and $\hat{h}_k(D)$, $\hat{k}_k(D)$ polynomials of degrees $n-1$ with piecewise constant coefficients; these polynomials are estimates of $h(D)$ and $k(D)$ in the fixed control strategy and they are defined by the respective entries in $\hat{\underline{\theta}}_k$. Definition (6.1) and Fig. 3 show the hybrid nature of the adaptive controller presented, as the feedback loop operates in continuous time while the parameters of the adaptive compensator ($\hat{h}_k(D)$ and $\hat{k}_k(D)$) are updated at discrete times t_k^0.

Also, since the systolic array estimation in the previous section provides the same estimated sequence $\{\hat{\underline{\theta}}_k\}$ as the recursive version (4.2), the same well-known stability arguments found in [6], [8] hold for this adaptive controller. Therefore we can conclude with the following theorem, which can be proved in a way analogous to [1] and [8].

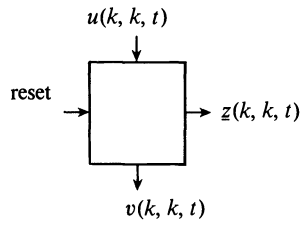
THEOREM. *The plant* (2.1) *with the estimation algorithm* (5.7), (5.8) *and the control input* (6.1) *is such that, for any uniformly bounded external input* $v(t)$ *the following holds with probability one*:

(i) *All signals and adaptive gains in the loop are uniformly bounded*;

(ii) $\lim_{t\to\infty} y(t) - y_m(t) = 0$ *for any initial conditions, with* $y_m(t)$ *the output of the reference model* (2.3).

7. Performance considerations and conclusions. The adaptive estimation using systolic arrays described in the previous sections is particularly suitable to block processing techniques as presented in [6] and [15], for systems where a large number of parameters

TABLE 1
Definitions of operations.



CELL (k, k)

IF reset = .NOT. TRUE
THEN
$c(k, t) = C(1, u(k, k, t-1))$
$s(k, t) = S(1, u(k, k, t-1))$

$$z(k,k,t) = \begin{bmatrix} c(k,t) \\ s(k,t) \end{bmatrix}$$

$v(k, k, t) = 0$

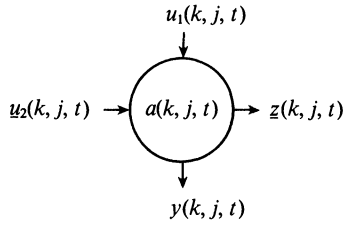
ELSE

$c(k, t) = 1$
$s(k, t) = 0$

$$z(k,k,t) = \begin{bmatrix} c(k,t) \\ s(k,t) \end{bmatrix}$$

$v(k, k, t) = u(k, k, t-1)$

ENDIF



$a(k, j, t) = F(u_1(k, j, t-1), a(k, j, t-1), \underline{u}_2(k, j, t-1))$
$y(k, j, t) = G(u_1(k, j, t-1), a(k, j, t-1), \underline{u}_2(k, j, t-1))$
$\underline{z}(k, j, t) = \underline{u}_2(k, j, t-1)$

Definitions

$C(a, u) = u/(u^2 + a^2)$; $S(a, u) = a/(u^2 + a^2)$

$$F(u_1, a, \underline{u}_2) = \underline{u}_2^T \begin{bmatrix} u_1 \\ a \end{bmatrix}$$

$$G(u_1, a, \underline{u}_2) = \underline{u}_2^T \begin{bmatrix} a \\ -u_1 \end{bmatrix}$$

has to be estimated. In particular the structure of the adaptive controller for multivariable systems [1], [17] is identical to the one presented above, with the difference that the parameter vector $\underline{\theta}^*$ encloses all the parameters of the multivariable compensator. It is

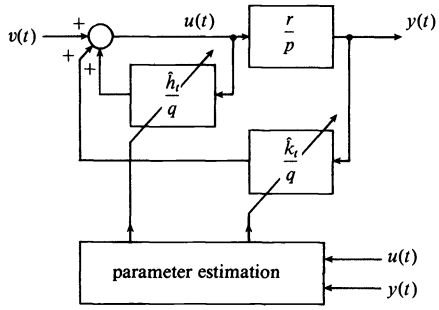


FIG. 3. *Direct adaptive control.*

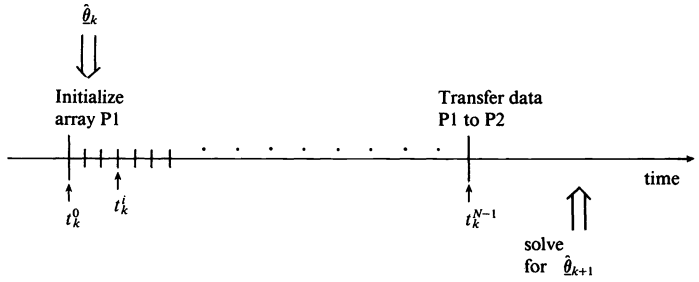


FIG. 4

easy to imagine how the complexity of the computations to be executed on line increases with the order of the system and its number of inputs and outputs.

In a standard RLS implementation as in (4.2) the sampling interval T in the sequence (3.1) must take the computation time into account, in order to guarantee a steady throughput of information without accumulation of data. It is clear that for the RLS algorithm (3.1) the sampling time interval T cannot be smaller than the time it takes to compute the new estimate, which is of the order $O(n^2)$ due to the required on line matrix manipulations. On the other hand because of the very nature of the systolic array structure, new data $\underline{\phi}(t_k^i)$ can enter the array right after each cell has performed its local computation, which is independent of the complexity of the system (its order n). Therefore the sampling time T is of the order $T = O(1)$.

The time-consuming part of the parallel implementation occurs at the end of the time block, when the data from the array (i.e., the triangular matrix $R_k(N-1)$ and $\beta_k(N-1)$) are transferred to the processor P2 and the linear system (4.8) is solved. This time delay, which occurs only once for each time block, increases linearly with the plant complexity n, as $T_d = O(n)$. Figure 4 summarizes these considerations.

REFERENCES

[1] G. C. GOODWIN, P. J. RAMADGE, AND P. E. CAINES, *Discrete time multivariable adaptive control*, IEEE Trans. Automat. Control, V AC 25 (1980), pp. 449–456.

[2] K. S. NARENDRA, Y. H. LIN, AND L. S. VALAVANI, *Stable adaptive controller design—Part* II: *Proof of stability*, IEEE Trans. Automat. Control, V AC 25 (1980), pp. 440–449.

[3] K. S. NARENDRA AND A. M. ANNASWAMY, *Robust adaptive control in the presence of bounded disturbances*, Tech. Report, Center for Systems Science, Yale University, New Haven, CT, 1985.

[4] P. A. IOANNOU AND K. TSAKALIS, *A robust adaptive controller*, Tech. Report, Electrical Engineering Systems, University of Southern California, Los Angeles, CA, 1985.

[5] B. D. O. ANDERSON AND R. M. JOHNSTONE, *Global adaptive pole positioning*, IEEE Trans. Automat. Control, V AC 30 (1985), pp. 11–21.

[6] H. ELLIOTT, R. CRISTI, AND M. DAS, *Global stability of adaptive pole placement algorithms*, IEEE Trans. Automat. Control, V AC 30 (1985), pp. 348–357.

[7] G. KREISSELMEIER, *Adaptive control via adaptive observation and asymptotic feedback matrix synthesis*, IEEE Trans. Automat. Control, V AC 25 (1980), pp. 717–722.

[8] G. C. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[9] B. D. O. ANDERSON AND C. R. JOHNSON, JR., *Exponential convergence of adaptive identification and control algorithms*, Automatica, 18 (1982), pp. 1–15.

[10] S. Y. KUNG AND J. T. JOHL, *VLSI wavefront arrays for image processing*, in VLSI for Pattern Recognition and Image Processing by K. S. Fu (ed.), Springer-Verlag, New York, Berlin, 1984.

[11] H. T. KUNG AND C. E. LEISERSON, *Systolic Arrays (for VLSI)*, Sparse Matrix Proceedings 1978, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[12] H. ELLIOTT, *Hybrid adaptive control of continuous time systems*, IEEE Trans. Automat. Control, AC 27 (1982), pp. 419–426.

[13] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[14] L. LJUNG AND T. SODERSTROM, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.

[15] G. C. GOODWIN, E. K. TEOH, AND B. C. MCINNIS, *Globally convergence adaptive controllers for linear systems with arbitrary zeroes*, Tech. Report, University of Newcastle, New South Wales, Australia, May 1982.

[16] W. GIVENS, *Computation of plane unitary rotations transforming a general matrix to triangular form*, J. Society Industrial Applied Math., (1958), pp. 26–50.

[17] H. ELLIOTT, W. WOLOVICH, AND M. DAS, *Arbitrary adaptive pole placement for multivariable systems*, IEEE Trans. Automat. Control, V AC 29 (1984), pp. 221–229.

[18] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 1, January 1988

SENSITIVITY ANALYSIS OF DIGITAL FILTER STRUCTURES*

VICTOR E. DEBRUNNER† AND A. A. (LOUIS) BEEX†

Abstract. A reasonable coefficient sensitivity measure for state space, recursive, finite wordlength, digital filters is the sum of the L_2 norm of all first-order partial derivatives of the system function with respect to the system parameters. This measure is actually a linear lower bound approximation to the output quantization noise power. An important feature of this measure is that it can be broken down into evaluations of ARMA auto- and cross-covariance sequences, both of which can be computed efficiently and in closed form. This efficient closed form computation is a big improvement over the computational methods used by previous researchers. Their limited methods produced only approximations to the sensitivity measure and wasted computer time (i.e., these methods are open form solutions). The direct form II sensitivity, which is shown to be approximately inversely proportional to the sum of products of system pole and zero distances, can, as a result, usually be reduced by the judicious placement of added pole/zero cancellation pairs. These cancellation pairs provide extra degrees of freedom which are used to minimize the sensitivity measure while not affecting the system function. This new filter still has the convenient direct form II structure.

Key words. sensitivity, L_2 norm, ARMA covariances, pole/zero cancellation pairs, quantization noise power

AMS(MOS) subject classification. 94C99

1. Introduction.

Much effort has recently been concentrated on the development of state space, recursive digital filters with low, or minimum, output quantization power. Jackson [12], [13], Kawamata and Higuchi [17], Tavsanoglu and Thiele [28], and Rao [26] have all examined the relationship between the coefficient sensitivity and the output quantization noise power; the sensitivity measure is a linear lower bound to the nonlinear output quantization noise power. Mullis and Roberts [22] and Hwang [11] developed, by different methods, the theoretical aspects of minimum noise filters as well as the practical computation of this optimal form. Recognizing that this optimal form has in general a full state space description, Mullis and Roberts developed a block-optimal form which is near optimal but has only $4n$ coefficients instead of the $n(n+2)$ coefficients of the optimal form. Later, Jackson, Lindgren, and Kim [14] developed a set of design equations for optimal second-order sections. Easily computed, this section-optimal form is identical to the block-optimal form above for parallel subfilters, while less optimal for cascaded subfilters. Continuing this process of eliminating coefficient count at the expense of added output quantization noise, Bomar and Hung [2] and Bomar [3], [4] have developed near-optimal second-order structures with constraints placed on the coefficient values so that some become structural ones and zeros while others become exact powers of two, thus making multiplications become simply shifts of the binary point. This above form is still less optimal than even the section-optimal form.

From a differing viewpoint, several researchers have devised design methods that use structures with known low output quantization noise power as the basic building blocks for the desired filter function. Among these building blocks are the wave digital filters of Fettweis [9] and their special case, the wave lattice digital filters [10]. Constantinides and Valenzuela [5], [6] noted the applicability of using all-pass functions to implement these filter types. Then, realizing the low output quantization noise power in

* Received by the editors June 20, 1987; accepted for publication October 1, 1987. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† Bradley Department of Electrical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

the pass-band, Vaidyanathan, Mitra, and Neuvo [29] developed a synthesis approach suitable for the design of low-pass filters which have low quantization noise power in the filter pass-band frequencies; however, the stop-band frequencies may have large noise powers, thus creating large output quantization noise power problems.

We look here at a completely new approach to the design of low output quantization noise power filters which do not have too many added filter coefficients; we increase the system order of canonical direct form II structures (as described in [24], [19], [15], [25], [27]) so as to reduce the output quantization noise power and not to change the system transfer function. We use the direct form II filter form because it is easy to compute, because it requires only $2n$ coefficients to implement, and because the sensitivity measure can be easily examined and simply reduced. In the process, we introduce the sensitivity measure and cursorily examine its validity by deriving it from both deterministic and stochastic system viewpoints. We also show a computational method which is both simple and very efficient, a marked improvement over earlier procedures.

2. The sensitivity measure. The sensitivity minimization will be based on the following state space description for a digital filter with impulse response h_n and rational transfer function $H(z)$:

$$x_{k+1} = Ax_k + Bu_k, \tag{1}$$

$$y_k = Cx_k + du_k \tag{2}$$

where x is the state vector, u is the input, and y is the output. Note that A is an $(n \times n)$ matrix, B is an $(n \times 1)$ vector, C is a $(1 \times n)$ vector, and d is a scalar. Further, the system transfer function is

$$H(z) = C(zI - A)^{-1}B + d. \tag{3}$$

Also, the state space representation $\{A, B, C, d\}$ is not unique. For any nonsingular $(n \times n)$ matrix T, the system has the algebraically equivalent state space description $\{T^{-1}AT, T^{-1}B, CT, d\}$.

Two different interpretations, one deterministic and the other probabilistic, exist for determining the sensitivity measure. In the deterministic view, the classic linearization procedure is used to approximate the nonlinear quantization effects. In the probabilistic view, the nonlinear quantization effects are modeled by injected noise sources. Both of these interpretations have merit and since they both generate the same final sensitivity measure, they lend credence to each other.

First, we examine the deterministic case. The filter $H(z)$ is a function of the parameter set $\gamma = [\gamma_1, \gamma_2, \cdots, \gamma_l]$, where both l and γ depend on the particular implementation used. The set γ is the quantization of the set γ_∞, which is the set of ideal coefficients. If we expand the filter using a Taylor series around the ideal filter, the actual filter $H(z)$ that is implemented can be represented, as in Fig. 1, by the parallel combination of the ideal transfer function $H_\infty(z)$ described by γ_∞ and the error or stray transfer function $H_{\text{stray}}(z)$. Considering only the first-order terms by truncating the higher-order terms of $H_{\text{stray}}(z)$ (i.e., linearizing around the ideal transfer function) gives

$$\begin{aligned} H(z) &\cong H_\infty(z) + \left.\frac{\partial H^t(z)}{\partial \gamma}\right|_{\gamma_\infty} \delta\gamma \\ &\cong H_\infty(z) + \frac{\partial H^t(z;\gamma_\infty)}{\partial \gamma}\delta\gamma \end{aligned} \tag{4}$$

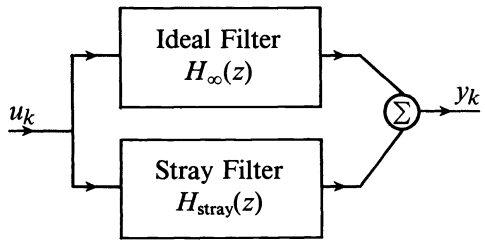


FIG. 1. *The linearized system.*

where

$$\frac{\partial H^t(z)}{\partial \gamma} = \left[\frac{\partial H(z)}{\partial \gamma_1}, \frac{\partial H(z)}{\partial \gamma_2}, \cdots, \frac{\partial H(z)}{\partial \gamma_l}\right]. \tag{5}$$

The L_2 norm, as described first by Tavsanoglu and Thiele [28] and later by Rao [26], yields a sensitivity measure; the square of the L_2 norm of the error (stray filter) is given by[1]

$$\frac{1}{2\pi j}\int \left|\frac{\partial H^t(z;\gamma_\infty)}{\partial \gamma}\delta\gamma\right|^2 \frac{dz}{z} \leqq \|\delta\gamma\|_2^2 \sum_{i=1}^{l} \frac{1}{2\pi j}\int \left|\frac{\partial H(z;\gamma_\infty)}{\partial \gamma_i}\right|^2 \frac{dz}{z}. \tag{6}$$

From the probabilistic viewpoint, the exact nature of the quantization effects is uncertain, which leads to the statistical model of Fig. 2. Note that the quantized branch is modeled as the ideal branch with a quantization noise term added. This added quantization noise is such that for the same input signal both branch models have the same output signal. The quantization noise terms are modeled using the following standard assumptions [24]:

(1) The sequence $\{\delta\gamma_{i,n}\}$ is a white noise process.

(2) The error sequences are uncorrelated with the other error sequences.

(3) The error sequences are uncorrelated with the input ν_n.

(4) The probability density function of the error process is uniform over the range of quantization error.

These assumptions lead to a linear probabilistic model for coefficient quantization. Heuristically, the model is valid when the input signal is sufficiently complex and the quantization steps are sufficiently small so that the amplitude of the input signal is likely to traverse many quantization levels from sample to sample. This model is supported empirically [24], where speech signals quantized to as low as eight bits exhibited these properties. The use of the above probabilistic model leads to the following state space descriptions for the effect of quantizing single parameter branches:

$$H(z) = (C + \delta c_i e_i^t)(zI - A)^{-1}B, \tag{7a}$$

$$H(z) = C(zI - A)^{-1}(B + \delta b_i e_i), \tag{7b}$$

$$H(z) = C(zI - (A + \delta a_{ij} e_i e_j^t))^{-1}B \tag{7c}$$

where e_i is the unit length vector with a one in the ith position and zeros elsewhere. Note that assumptions 2 and 3 above allow the separation of the errors as described in (7).

[1] Note that in this and all other integrations in this work, $\int$ denotes contour integration along the unit circle of the z-plane in the counterclockwise direction.

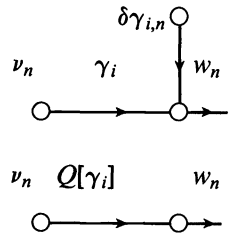


FIG. 2. *The probabilistic model.*

Clearly, the coefficient quantization errors in the C vector (7a) are propagated through the system function as

$$\delta c_i e_i^t (zI-A)^{-1}B = \delta c_i \frac{\partial H(z)}{\partial c_i} \tag{8}$$

(see (16)) while the coefficient quantization errors in the B vector (7b) are propagated through the system as

$$C(zI-A)^{-1}\delta b_i e_i = \delta b_i \frac{\partial H(z)}{\partial b_i} \tag{9}$$

(see (15)). To separate the coefficient quantization errors in the A matrix, we use the Sherman–Morrison formula [30]

$$[(zI-A)-e_i e_j^t \delta a_{ij}]^{-1} = (zI-A)^{-1} + \frac{(zI-A)^{-1}e_i e_j^t (zI-A)^{-1}\delta a_{ij}}{1-e_j^t(zI-A)^{-1}e_i \delta a_{ij}}.$$

Thus the output error transfer function is given by

$$\frac{C(zI-A)^{-1}e_i e_j^t (zI-A)^{-1}B\delta a_{ij}}{1-e_j^t(zI-A)^{-1}e_i\delta a_{ij}}.$$

By the assumptions on the quantization, the denominator is very close to one; thus the error term is approximately given by

$$C(zI-A)^{-1}e_i e_j^t (zI-A)^{-1}B\delta a_{ij} = \delta a_{ij}\frac{\partial H(z)}{\partial a_{ij}} \tag{10}$$

(see (17)). Thus, we finally can describe the system as in Fig. 3.

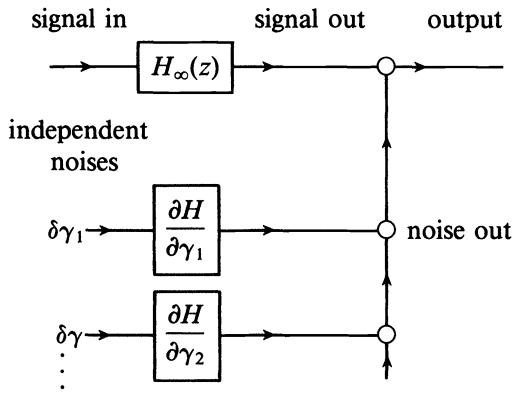


FIG. 3. *The probabilistic system.*

Taking the mean square value of the error (output noise) terms gives

$$E\left[\frac{1}{2\pi j}\int\left|\frac{\partial H'(z;\gamma_\infty)}{\partial\gamma}\delta\gamma\right|^2\frac{dz}{z}\right]=\sigma_o^2\sum_{i=1}^{l}\frac{1}{2\pi j}\int\left|\frac{\partial H(z;\gamma_\infty)}{\partial\gamma_i}\right|^2\frac{dz}{z} \tag{11}$$

where σ_o^2 is the noise variance of a single quantizer of the system. Since the quantization assumed is rounding, $E[\delta\gamma_i] = 0$ and the variance is given by

$$\sigma_o^2=\frac{2^{-2b}}{12} \tag{12}$$

where b is the coefficient wordlength in bits. This probabilistic criterion has been used by several researchers [18] to quantify the transfer function degradation caused by finite wordlength effects.

Noting the similarity between (6) and (11), Rao [26] defined the L_2 norm sensitivity measure S_2 as

$$\begin{aligned}S_2&\equiv\sum_{i=1}^{l}\frac{1}{2\pi j}\int\frac{\partial H(z)}{\partial\gamma_i}\frac{\partial H(z^{-1})}{\partial\gamma_i}\frac{dz}{z}\\&=\sum_{i=1}^{l}\frac{1}{2\pi j}\int\left|\frac{\partial H(z)}{\partial\gamma_i}\right|^2\frac{dz}{z}\\&=\sum_{i=1}^{l}\left\|\frac{\partial H(z)}{\partial\gamma_i}\right\|_2^2\end{aligned} \tag{13}$$

where the γ_i are the nonstructural coefficients (i.e., the coefficients that are $\not\equiv 0$ or $\not\equiv \pm 1$) of the $\{A, B, C\}$ state space description.

For further justification of using the S_2 measure as an indication of output quantization noise power, Jackson [12] has derived roundoff noise bounds from these coefficient sensitivities. Of special interest is the lower bound

$$\sigma_o^2 S_2 \leqq \sigma_e^2 \tag{14}$$

where σ_e^2 is the filter output quantization noise variance. That S_2 is a lower bound for σ_e^2 is also evident from Fig. 1, remembering that S_2 is the output power of a truncated form of the stray transfer function. Calculating the output variance of $H_{\text{stray}}(z)$ (with all the terms present) gives an infinite sum of auto-covariance terms because all the cross terms go to zero under the assumption that the quantization noise sources are statistically independent from each other and the input signal source. Remember that S_2 is only one of these auto-covariances, although it will be the largest one because of the order. The lower bound of (14) was shown empirically by Jackson to be a rather tight bound; thus S_2 is closely related to the output noise power (data presented in the example section confirm the boundedness). Since one number, i.e., the coefficient sensitivity measure S_2, describes the filter quantization noise power, the problem of identifying low roundoff noise filters is made conceptually easy.

3. Computing the sensitivity measure. In this section we concern ourselves with calculating the sensitivity measure S_2. First we determine the necessary partial derivatives. The partial derivatives, with respect to b_i and c_i, of $H(z)$ in (3) lead directly to the first-order sensitivity functions:

$$\frac{\partial H(z)}{\partial b_i}=C(zI-A)^{-1}e_i, \tag{15}$$

$$\frac{\partial H(z)}{\partial c_i}=e_i^t(zI-A)^{-1}B. \tag{16}$$

Somewhat more difficult to determine are the sensitivity functions for the coefficients of the A matrix. Using the mathematical identity

$$\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1}\frac{\partial A}{\partial \alpha}A^{-1}$$

gives

$$\frac{\partial H(z)}{\partial a_{ij}} = -C(zI-A)^{-1}\frac{\partial (zI-A)}{\partial a_{ij}}(zI-A)^{-1}B$$

or

$$\frac{\partial H(z)}{\partial a_{ij}} = -C(zI-A)^{-1}e_i e_j^t (zI-A)^{-1}B$$

which is simply

$$\frac{\partial H(z)}{\partial a_{ij}} = \frac{\partial H(z)}{\partial b_i}\frac{\partial H(z)}{\partial c_j}. \tag{17}$$

Notice that these sensitivity functions are rational functions with the same poles as the original transfer function, $H(z)$; thus, the unit circle is in the region of convergence of the integrand. Evaluation of these partial derivatives is simply a matter of computing transfer functions from the state space description. We will not discuss this well-known problem, but merely refer the reader to Melsa [20] for an effective comment on this subject.

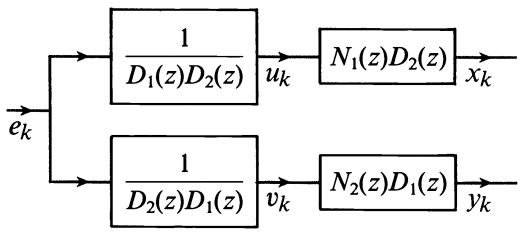
The complex integration necessary in computing the sensitivity measure is recognized to be a covariance of the ARMA system H_1 and H_2 of Fig. 4 at lag $k = 0$ with the system input being zero mean white noise; i.e.,

$$r_{xy}(k) = \frac{1}{2\pi j}\int H_1(z)H_2(z^{-1})z^{k-1}\,dz.$$

This integration is performed using an algorithm for the calculation of ARMA auto- and cross-covariances presented by Dugre, Beex, and Scharf [8] and by Beex [1], respectively. Note that an auto-covariance is generated when $H_1(z) = H_2(z)$ in Fig. 4 and that we are then not required to imbed the polynomials in step 1 of the general cross-covariance algorithm given below since $D_1(z) = D_2(z)$. The cross-covariance algorithm requires four steps.

(1) Imbed the polynomials

$$\tilde{D}_1(z) = D_1(z)D_2(z), \qquad \tilde{D}_2(z) = D_1(z)D_2(z),$$
$$\tilde{N}_1(z) = N_1(z)D_2(z), \qquad \tilde{N}_2(z) = D_1(z)N_2(z).$$



$$H_1(z) = \frac{N_1(z)}{D_1(z)} \qquad H_2(z) = \frac{N_2(z)}{D_2(z)}$$

FIG. 4. *The covariance generator system.*

(2) With $\tilde{D}_1(z)$, use scalar Levinson recursion to generate the auto-covariance of the AR part, checking the magnitude of the reflection coefficients to determine system stability and thus covariance generation sensibility.

(3) From $\tilde{N}_2(z)$ and $\tilde{N}_1(z)$ determine $\tilde{F}$ from the convolution

$$\tilde{f}_k = (\tilde{n}_2)_{-k} * (\tilde{n}_1)_k$$

where the $\tilde{n}_2$ are the coefficients of $\tilde{N}_2(z)$ and the $\tilde{N}_1(z)$ and the $\tilde{f}_k$ are similarly defined.

(4) Convolve the AR auto-covariance with $\tilde{f}_k$ to get the ARMA cross-covariance sequence r_k, $k = 0, 1, 2, \cdots, n$.

The value of the required contour integral is then equal to r_0. This algorithm is simple to implement on the computer and yields good numerical results except where noted in § 5.

Since all the partial derivatives are rational functions with the same poles as the original function, the unit circle is in the region of convergence of the integrand (assuming, of course, that the system is stable). Hence, the sensitivity measure may be viewed as the sum of the variances of stable, rational ARMA systems. We have just shown that we can compute these variances efficiently and in closed form, a marked improvement over the three computational methods which previous researchers used. As will be clear, we have achieved both accuracy and computational efficiency and gained a computational generality previously unattained. The following is a discussion of these methods.

(1) Tavsanoglu and Thiele [28] defined

$$S_A(z) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial H(z)}{\partial a_{ij}}, \qquad a_{ij} \not\equiv 0, \pm 1,$$

$$S_B(z) = \sum_{i=1}^{n} \frac{\partial H(z)}{\partial b_i}, \qquad b_i \not\equiv 0, \pm 1, \tag{18}$$

$$S_C(z) = \sum_{j=1}^{n} \frac{\partial H(z)}{\partial c_j}, \qquad c_j \not\equiv 0, \pm 1$$

and using these relations, defined the sensitivity measure as

$$S_2' = \|S_B\|_2^2 \cdot \|S_C\|_2^2 + \|S_B\|_2^2 + \|S_C\|_2^2. \tag{19}$$

Clearly, this measure is an upper bound approximation to our S_2 given by (13), seen by using (17) and the Cauchy–Schwartz inequality to give an upper bound to $\|S_A\|_2^2$, i.e.,

$$\|S_A\|_2^2 = \|S_B S_C\|_2^2 \leqq \|S_B\|_2^2 \cdot \|S_C\|_2^2.$$

Now, Tavsanoglu and Thiele compute their upper bound approximation to S_2 by noting that [22], [28]

$$\|S_B\|_2^2 = \sum_{i=1}^{n} w_{ii} = \operatorname{tr} W, \qquad b_i \not\equiv 0, \pm 1 \tag{20}$$

and

$$\|S_C\|_2^2 = \sum_{j=1}^{n} k_{jj} = \operatorname{tr} K, \qquad c_j \not\equiv 0, \pm 1 \tag{21}$$

where W and K are the observability and controllability grammians, respectively. These grammians are the solutions to the Lyapunov equations

$$K = AKA^t + BB^t, \qquad W = A^t WA + C^t C. \tag{22}$$

It is well known that the solutions are the infinite sums

$$K = \sum_{i=0}^{\infty} A^i B B^t (A^t)^i, \qquad W = \sum_{j=0}^{\infty} (A^t)^j C^t C A^j. \tag{23}$$

Truncating these infinite sums at some finite i, j gives an approximation of K and W to a given accuracy and thus gives an approximation to the upper bound S_2'. These computations are time-consuming; also their accuracy is circumspect [22]. Furthermore, we can only compute an upper bound, because of the simplification in (19).

(2) Rao [26] takes a slightly different approach to the computational problem. He notes that

$$\|S_A\|_2^2 = \|S_B\|_2^2 \cdot \|S_C\|_2^2 + 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{p=1}^{\infty} W_i^t A^p e_i e_j^t A^p K_j, \qquad a_{ij} \neq 0, \pm 1 \tag{24}$$

where W_i and K_j are the ith and jth columns of W and K, respectively. Thus, S_2 may be computed exactly by

$$S_2 = \|S_B\|_2^2 \cdot \|S_C\|_2^2 + \|S_B\|_2^2 + \|S_C\|_2^2 + 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{p=1}^{\infty} W_i^t A^p e_i e_j^t A^p K_j. \tag{25}$$

As in Tavsanoglu and Thiele's computational method, we must approximate infinite sums by finite ones.

(3) Alternatively, Knowles and Olcayto [18] suggest the less elegant method shown in Fig. 5 to calculate the variances necessary in computing the sensitivity measure. Again, however, we must evaluate approximations to infinite sums, and this evaluation is both wasteful of computer time and only as accurate as the approximation allows.

Our method of calculation is both more efficient and more accurate than the above three methods since it is a closed-form method.

To digress a moment, it is clear from the above discussion that the controllability and observability grammians may also be solved in closed form. The controllability grammian K may be computed as

$$k_{ij} = \frac{1}{2\pi j} \int \frac{\partial H(z)}{\partial c_i} \frac{\partial H(z^{-1})}{\partial c_j} \frac{dz}{z}, \qquad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, i. \tag{26a}$$

Since K is symmetric, we need only calculate $\frac{1}{2}n(n+1)$ terms, not the n^2 elements as would otherwise be needed. Thus, the controllability grammian can be efficiently evaluated via a closed-form, exact procedure.

(1) If $i = j$, then perform an auto-covariance computation with the ARMA covariance generator.

(2) If $i \neq j$, then perform a cross-covariance computation with the ARMA covariance generator.

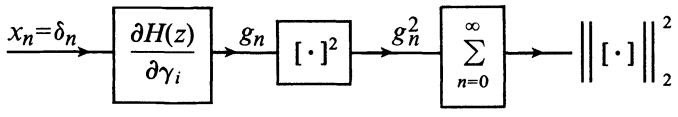


FIG. 5. *Variance generator used by Knowles and Olcayto.*

Of later importance, observe that the controllability grammian for a direct form II state space filter is the covariance matrix

$$K=\begin{bmatrix} r_0 & r_1 & r_2 & \cdots & r_{n-1} \\ r_1 & r_0 & r_1 & \cdots & r_{n-2} \\ & & \vdots & & \\ r_{n-1} & r_{n-2} & \cdots & r_1 & r_0 \end{bmatrix}$$

where

$$r_k=\frac{1}{2\pi j}\int \frac{\partial H(z)}{\partial c_i}\frac{\partial H(z^{-1})}{\partial c_i}z^{k-1}\,dz, \qquad k=0,1,\cdots,n-1 \tag{26b}$$

and i is any valid subscript. The r_k are the auto-covariance sequence members of the partial derivative system in (26b) and may thus easily be computed in exact closed form by our ARMA covariance generator. Similar to the controllability grammian, the observability grammian W can be computed as

$$w_{ij}=\frac{1}{2\pi j}\int \frac{\partial H(z)}{\partial b_i}\frac{\partial H(z^{-1})}{\partial b_j}\frac{dz}{z}, \qquad i=1,2,\cdots,n, \quad j=1,2,\cdots,i. \tag{27}$$

Since W is symmetric, we need only calculate $\frac{1}{2}n(n+1)$ terms, not the n^2 elements as would otherwise be needed. As in the controllability grammian, the observability grammian can be efficiently evaluated via a closed-form, exact procedure.

(1) If $i = j$, then perform an auto-covariance computation with the ARMA covariance generator.

(2) If $i \neq j$, then perform a cross-covariance computation with the ARMA covariance generator.

Thus, we can efficiently calculate both grammians K and W in closed form.

4. Output quantization error. For completeness, it is necessary to compute an estimate for the output quantization noise power; this calculation corroborates the S_2 measure as shown in (14). Hence the calculation of the roundoff noise power is important for verification purposes. Note that we are calculating the true output noise variance of the error transfer function, $H_{\text{stray}}(z)$, which was described in § 2.

4.1. The error state space description. Under finite wordlength conditions, the elements of A, B, and C, as well as the scalar d, are constrained and the corresponding state space representation becomes

$$\hat{x}_{k+1}=\hat{A}\hat{x}_k+\hat{B}u_k, \tag{28}$$

$$\hat{y}_k=\hat{C}\hat{x}_k+\hat{d}u_k \tag{29}$$

where $\hat{}$ denotes a quantized entity. The quantization of the input u_k is ignored since we are studying system generated quantization noise only. Thus the error $e_k = y_k - \hat{y}_k$ is the difference between the ideal (infinite wordlength) output y_k and the quantized (finite wordlength) output $\hat{y}_k$.

The error state space filter can be constructed as follows:

$$e_k=y_k-\hat{y}_k=Cx_k-\hat{C}\hat{x}_k+(d-\hat{d})u_k$$

or,

$$e_k=[C,-\hat{C}]\begin{bmatrix} x_k \\ \hat{x}_k \end{bmatrix}+(d-\hat{d})u_k \tag{30}$$

with a corresponding system equation:

$$\begin{bmatrix} x_{k+1} \\ \hat{x}_{k+1} \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \hat{x}_{k+1} \end{bmatrix} + \begin{bmatrix} B \\ \hat{B} \end{bmatrix} u_k. \tag{31}$$

Consequently, the error transfer function $H_e(z)$ is given by

$$H_e(z) = [C, -\hat{C}] \left\{ zI - \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} \right\}^{-1} \begin{bmatrix} B \\ \hat{B} \end{bmatrix} + (d - \hat{d}). \tag{32}$$

Now the output error variance is

$$\sigma_e^2 = \frac{1}{2\pi j} \int H_e(z) H_e(z^{-1}) \frac{dz}{z}. \tag{33}$$

Clearly, the quantizations that occur when forming $\hat{A}$, $\hat{B}$, $\hat{C}$, and $\hat{d}$ in (32) depend on the form of the state space realization (i.e., on the form of the A, B, and C matrices). Furthermore, the quantizations determine the exact form of $H_e(z)$. Thus, the state space realization affects σ_e^2 in (33) and we expect to be able to classify filter realizations that minimize the output noise power.

4.2. Block diagram view of output noise. Alternatively, we may view the output error as the difference in output of $H(z)$ and $\hat{H}(z)$ when driven by the same input (see Fig. 6). The output noise variance, $E\{e_k^2\} \equiv \sigma_e^2$, can be readily found as follows:

$$\begin{aligned} E\{e_k^2\} &= E\{(y_k - \hat{y}_k)^2\} \\ &= E\{y_k^2\} + E\{\hat{y}_k^2\} - 2E\{\hat{y}_k y_k\} \end{aligned}$$

or,

$$\sigma_e^2 = \sigma_y^2 + \sigma_{\hat{y}}^2 - 2\sigma_{y\hat{y}}. \tag{34}$$

This computation is easily performed as follows:

$$\sigma_e^2 = \frac{1}{2\pi j} \int [H(z)H(z^{-1}) + \hat{H}(z)\hat{H}(z^{-1}) - 2H(z)\hat{H}(z^{-1})] \frac{dz}{z}. \tag{35}$$

Thus, σ_e^2 can alternatively be computed using the cross-correlation terms $\sigma_{y\hat{y}}$ together with the two auto-correlation terms σ_y^2 and $\sigma_{\hat{y}}^2$.

5. Practical computation notes. During the course of implementing and using the above sensitivity measures, two numerical problems in the cross-covariance generator were noted.

(1) Poles close to the unit circle may migrate to unstable positions outside the unit circle as a result of creating the higher-order (imbedded) polynomials in step 1 of the ARMA auto- and cross-covariance generator algorithm.

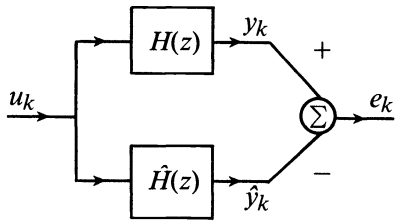


FIG. 6. *The error system block diagram.*

(2) Precision error in the last convolution of the algorithm (step 4) may actually produce a resulting negative auto-covariance, especially when large numbers are alternately added and subtracted!

The first problem may be eliminated by progressively increasing precision since the backward Levinson recursion of step 2 of the algorithm may generate large errors from small errors caused by the polynomial multiplication of step 1. This error was studied in detail by Cybenko [7]. The second problem is more difficult to anticipate and so, as discussed earlier, two methods to determine σ_e^2 were developed. The second method described by (35) appears to be numerically superior to that of (33), and so is generally preferred in calculating σ_e^2. This superiority was determined empirically from the example systems, but intuitively the reader should expect this superiority since the second method does not require a doubling of the system order. We note that these problems show up specifically when designing and analyzing filters approximating ideal characteristics in which poles are located almost on the unit circle.

6. Sensitivity of low-pass direct form II digital filters. A commonly used filter design technique is to determine the desired filter characteristics, then translate these characteristics to their corresponding low-pass filter equivalents, and finally design this normalized low-pass filter. The low-pass filter is then frequency transformed back to the desired type: either low-pass, band-pass, band-stop, or high-pass filter. Because of this practice, it is logical to first look at low-pass filters and then determine the characteristics related to their sensitivity measure which can be used to advantage.

6.1. Basic low-pass filter description. Of great interest to us is the fact that the poles are clustered near $z = 1$ inside the unit circle and have magnitudes close to one. The sensitivity measure S_2 of a direct form II filter will be shown to be approximately inversely proportional to the system pole distances. This is as follows [24]. Given the ideal system transfer function

$$H(z) = d + \frac{b_1 z^{-1} + b_2 z^{-2} + \cdots + b_m z^{-m}}{1 - a_1 z^{-1} - a_2 z^{-2} \cdots - a_n z^{-n}}, \qquad m \leqq n$$
$$= \frac{N(z)}{D(z)}. \tag{36}$$

Express the denominator, $D(z)$, as

$$D(z) = 1 - \sum_{j=1}^{n} a_j z^{-j} = \prod_{j=1}^{n} (1 - p_j z^{-1}) \tag{37}$$

where the p_j are the simple poles of $H(z)$. From calculus,

$$\left.\frac{\partial H(z)}{\partial p_i}\right|_{z=p_i} \frac{\partial p_i}{\partial a_j} = \left.\frac{\partial H(z)}{\partial a_j}\right|_{z=p_i} \tag{38}$$

which can be rewritten as

$$\frac{\partial p_i}{\partial a_j} = \left.\frac{\partial H(z)}{\partial a_j}\right|_{z=p_i} \Bigg/ \left.\frac{\partial H(z)}{\partial p_i}\right|_{z=p_i}$$
$$= \left.\frac{N(z)}{D(z)^2}\frac{\partial D(z)}{\partial a_j}\right|_{z=p_i} \Bigg/ \left.\frac{N(z)}{D(z)^2}\frac{\partial D(z)}{\partial p_i}\right|_{z=p_i} \tag{39}$$
$$= \left.\frac{\partial D(z)}{\partial a_j}\right|_{z=p_i} \Bigg/ \left.\frac{\partial D(z)}{\partial p_i}\right|_{z=p_i}.$$

Taking the required derivatives using (37), the pole sensitivity can be rewritten as

$$\frac{\partial p_i}{\partial a_j} = \frac{p_i^{n-j}}{\prod_{l \neq i}^{n} (p_i - p_l)}. \tag{40}$$

Similarly, the numerator $N(z)$ can be written

$$N(z) = \sum_{j=0}^{m} b_j z^{-j} = b_0 \prod_{j=1}^{m} (1 - z_j z^{-1}). \tag{41}$$

As for the denominator, the z_j are the simple zeros of $H(z)$. Parallel to (38), the zero sensitivity can be determined from

$$\left.\frac{\partial H(z)}{\partial z_i}\right|_{z=z_i} \frac{\partial z_i}{\partial b_j} = \left.\frac{\partial H(z)}{\partial b_j}\right|_{z=z_i} \tag{42}$$

which can be rewritten as

$$\begin{aligned} \frac{\partial z_i}{\partial b_j} &= \left.\frac{\partial H(z)}{\partial b_j}\right|_{z=z_i} \Big/ \left.\frac{\partial H(z)}{\partial z_i}\right|_{z=z_i} \\ &= \left.\frac{1}{D(z)} \frac{\partial N(z)}{\partial b_j}\right|_{z=z_i} \Big/ \left.\frac{1}{D(z)} \frac{\partial N(z)}{\partial z_i}\right|_{z=z_i} \\ &= \left.\frac{\partial N(z)}{\partial b_j}\right|_{z=z_i} \Big/ \left.\frac{\partial N(z)}{\partial z_i}\right|_{z=z_i} \end{aligned} \tag{43}$$

which from (41) reduces to

$$\frac{\partial z_i}{\partial b_j} = \frac{z_i^{m-j}}{\prod_{l \neq i}^{m} (z_i - z_l)}. \tag{44}$$

It is important for us to interpret this latter result in terms of the sensitivity measure S_2. From the definition of the S_2 sensitivity measure, an alternate way of writing S_2 for the direct form II state space is

$$S_2 = \frac{1}{2\pi j} \int \sum_{j=1}^{n} \left|\frac{\partial H(z)}{\partial a_j}\right|^2 + \sum_{i=1}^{m} \left|\frac{\partial H(z)}{\partial b_i}\right|^2 \frac{dz}{z}. \tag{45}$$

(The coefficients of the direct II form state space are coefficients of the system function $H(z)$.) Equations (39) and (43), along with (45), show that the S_2 sensitivity measure is proportional to the pole and zero sensitivities. Note that

$$\frac{\partial H(z)}{\partial a_k} = -\frac{N(z)}{D^2(z)} z^{-k} \tag{46}$$

which can be rewritten using (37) as

$$\frac{\partial H(z)}{\partial a_k} = -\frac{N(z)}{D(z)} \frac{z^{n-k}}{\prod_{j=1}^{n} (z - p_j)} \tag{47}$$

and similarly,

$$\begin{aligned} \frac{\partial H(z)}{\partial b_k} &= \frac{1}{D(z)} z^{-k} \\ &= \frac{N(z)}{D(z)} \frac{z^{n-k}}{\prod_{j=1}^{m} (z - z_j)}. \end{aligned} \tag{48}$$

Both (47) and (48) are similar to the pole and zero sensitivities of (40) and (44), with $H(z)$ as a weighting function. Further, the S_2 sensitivity measure is evaluated as the complex contour integral on the unit circle of the z-plane while the pole and zero sensitivities of (40) and (44) are only point evaluations at the pole and zero locations of the system (i.e., the sensitivity measure is an integration over all z on the unit circle). In practice this difference is of limited consequence, as the partial derivatives of the transfer function appear to approximate delta functions, thus making the integration itself close to a point evaluation. Since the pole and zero sensitivities are inversely proportional to the system pole and zero distance, the S_2 measure is also approximately inversely proportional to the system pole and zero distances. Since the sensitivity measure is weighted by the system transfer function, only that output quantization noise power in regions of practical importance (i.e., the noise power in frequencies passed by the filter) are considered.

Using (40), Kaiser [15], [16] showed that small errors in the coefficients can create large pole displacements from the ideal design. Coefficient quantization errors belong to the category of small errors, and so one would expect large S_2 sensitivities (and thus large quantization noise power) in narrow bandwidth low-pass filters, where the poles are tightly clustered.

6.2. Reducing direct form II low-pass filter sensitivities. Until recently, the principal method of reducing large roundoff noise power has been the classical analogue filter design approach of breaking large-order filters into cascaded or parallel second-order sections. Here, the complex conjugate poles are isolated from each other, and so the error in each pole is independent from its distance to all the other poles in the higher order system, thus reducing the overall system output quantization noise. However, forms have been developed that minimize the roundoff noise. The cost of this form is increased complexity; the transformation to the optimal form causes the $\{A, B, C\}$ state space description to be filled with nontrivial coefficients [22]. However, Mullis and Roberts [22] presented their block-optimal form and Jackson, Lindgren, and Kim [14] their section-optimal forms which have near-optimal output quantization noise power and reduced complexity.

Since the direct form II is trivial to compute (i.e., the state space coefficients are identical to the transfer function coefficients), the idea of reducing its sensitivity without altering its form is the attractive idea we pursue here. Equations (40) and (44) suggest a procedure for reducing the direct form II sensitivity by adding poles and zeros; because the transfer function must remain unchanged, the added pole must have a corresponding zero while any added zero should also have its identically related pole. In the case of low-pass filters, all the poles are at low frequencies and so are grouped near $z = 1$ in the z-plane. Clearly, if a pole/zero cancellation pair is added at a high frequency (near $z = -1$ in the z-plane), the sensitivity must be reduced because the added pole distances are greater than one. We also see that additional reductions in filter sensitivity can be achieved by adding a complex conjugate pair of pole/zero cancellations. For a comparison of the system complexity, note that an nth-order optimal form has $n(n + 2)$ nontrivial coefficients and the block-optimal form has $4n$ while the $(n + 2)$th-order (with two added degrees of freedom) direct form II has only $2(n + 2)$ nontrivial coefficients. Clearly, the reduced coefficient count begins to be beneficial for systems with order as low as two, and, as the order of the original system grows, the savings becomes an important issue. Problems associated with this method as pertaining to system order, system bandwidth, and system stability are best described in detail with their corresponding examples. At present it is sufficient to say that the method works best on small systems which do not have narrow bandwidths.

7. Examples. To illustrate the range of sensitivities for different implementations of the same system function, the third-order low-pass filter used by Hwang [11] was examined. The system has transfer function

$$H(z)=\frac{0.79306721z^{-1}+.023016947z^{-2}+.0231752363z^{-3}}{1-1.974861148z^{-1}+1.556161235z^{-2}-.4537681314z^{-3}}. \tag{49}$$

The forms and their sensitivities are as follows.

(1) The direct form II sensitivity measure is 93.714442.

(2) The cascade form sensitivity measure is 43.511076.

(3) The parallel form sensitivity measure is 15.698915.

(4) The optimal form (coefficients computed in [11]) sensitivity measure is 8.816327.

(5) The block-optimal form (coefficients computed according to the design equations given in [14]) sensitivity measure is 7.338480. Note that this is lower than the optimal form of Hwang in item 4 above, which may be due to not actually having the optimal form coefficients. In any case, the representation is nearly optimal.

(6) The section-optimal form (coefficients computed according to the design equations given in [14]) sensitivity measure is 24.787467.

The output quantization noise power estimates for the above implementations at various wordlengths are shown in Fig. 7 and the close relationship (refer to (14)) between S_2 and σ_e^2 is shown in Fig. 8 for the direct form II and the optimal form. Clearly, for this particular filter, all forms are relatively insensitive to coefficient quantization; even the direct form II is only an order of magnitude more sensitive than the optimal forms.

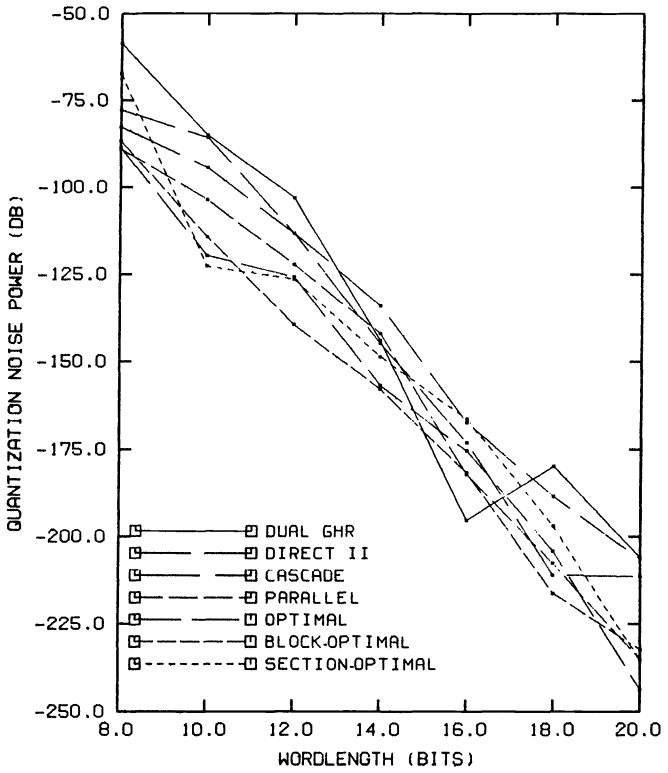


FIG. 7. *Output quantization noise power.*

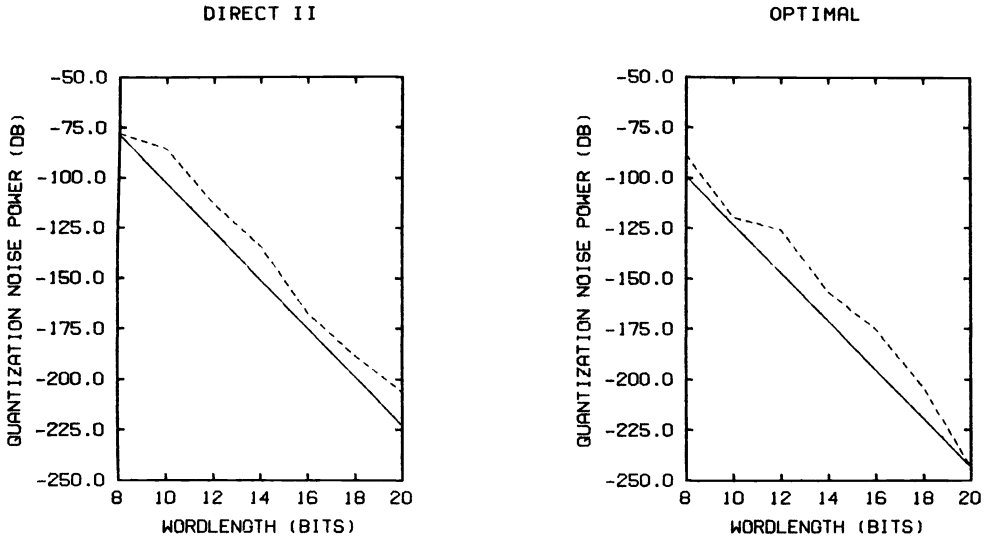


FIG. 8. *The sensitivity as a lower bound of the quantization noise power.*

To experiment, a single real pole/zero pair is added to the direct form II filter of (47). The sensitivity measure as a function of the location of a real pole/zero cancellation pair is given in Fig. 9, and it reveals a minimum sensitivity comparable to the sensitivity

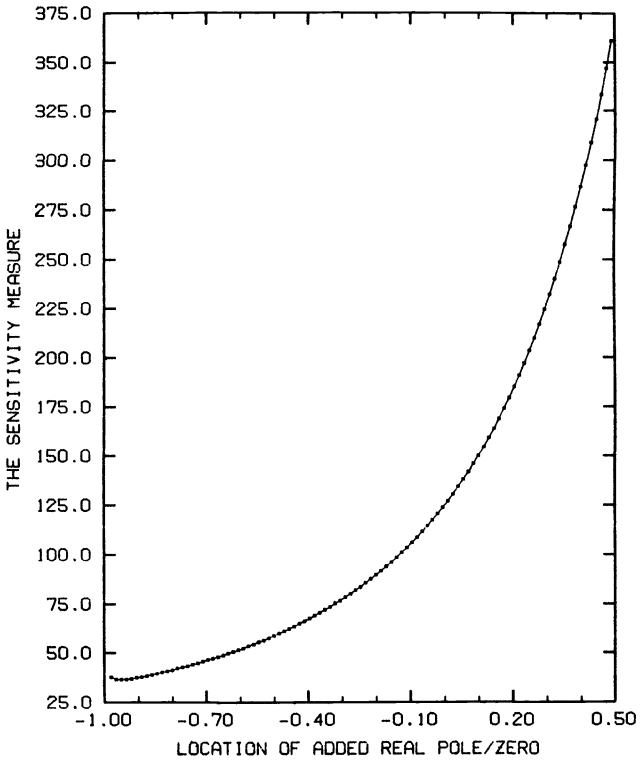


FIG. 9. *Sensitivity of fourth-order implementations.*

of the cascade realization, $S_2 = 36.601496$, with the single real pole/zero cancellation pair added at $z = -.952450$. The added pole/zero pair has increased the system order by one, thus adding two nontrivial coefficients above the number required for the original order direct form II model. The sensitivity of the filter has been reduced, but has the transfer function been changed in the process? Ideally, a pole/zero cancellation will leave the system transfer function unchanged, but without infinite precision wordlengths, the impulse response will change, however imperceptibly. This change, given in Fig. 10, shows that the system function has hardly been changed at all. As commented previously, further improvement in sensitivity can be realized when a complex conjugate pair of pole/zero cancellations is added.

From the sensitivity surface of Fig. 11, a location in the z-plane is found which has a sensitivity lower than the minimum sensitivity achieved by adding only a single, real pole/zero cancellation pair. A sensitivity, $S_2 = 18.562108$, is obtained when a pair of pole/zero cancellations are added at radius $r = .908248$ and angles $\theta = \pm 150.981$ degrees in the z-plane. This sensitivity is comparable to the sensitivity of the parallel description, while not quite twice as sensitive as the optimal form. Again, the cost is not too great since only four nontrivial coefficients are added. Also, as before, the transfer function has only changed to the same extent as above.

To summarize the improvements, Fig. 12 compares the output quantization noise power of the optimal and various direct form II implementations of the third-order filter. Note that the higher-order, reduced sensitivity direct form II have lower output quantization noise power at every bit wordlength than does the third-order original direct form II filter. Also, both of these direct form II filters actually have lower output quantization noise power than the optimal form at certain wordlengths.

To show that the added pole/zero cancellation approach will reduce the sensitivity for larger-order systems as well, a new example filter is introduced. Larger sensitivity reductions are expected because of the pole placements (and thus the pole distances), but the direct form II sensitivity will also be intrinsically much higher because of the higher number of poles and corresponding pole distances which are much less than one. The filter is a tenth-order all-pole low-pass function with system transfer function

$$H(z) = \frac{N(z)}{D(z)} \tag{50}$$

where

$$N(z) = .211348904z^{-1},$$

$$\begin{aligned} D(z) = 1 &- 5.24714092z^{-1} + 14.6742367z^{-2} - 27.2976798z^{-3} \\ &+ 37.1004172z^{-4} - 38.082725z^{-5} + 29.9060915z^{-6} \\ &- 17.7209547z^{-7} + 7.66182077z^{-8} - 2.20028154z^{-9} + .339082688z^{-10}. \end{aligned}$$

The direct form II has $S_2 = 2{,}109{,}022{,}068.714$. Placing a complex conjugate pair of pole/zero cancellations at radius $r = 0.99$ and angle $\theta = \pm 180$ degrees in the z-plane reduces the sensitivity measure to $S_2 = 199{,}434{,}498.555$. Clearly we have reduced the sensitivity, but the filter still remains inordinately sensitive.

For narrow-bandwidth low-pass filters, the coefficient sensitivity can also be reduced using this method; however, because coefficient sensitivities of direct form II, as well as

cascade and parallel, implementations increase as bandwidths decrease under frequency transformations and the optimal form sensitivity is invariant to frequency transformations (Mullis and Roberts [23] and Kawamata and Higuchi [17]), the reduced sensitivity does

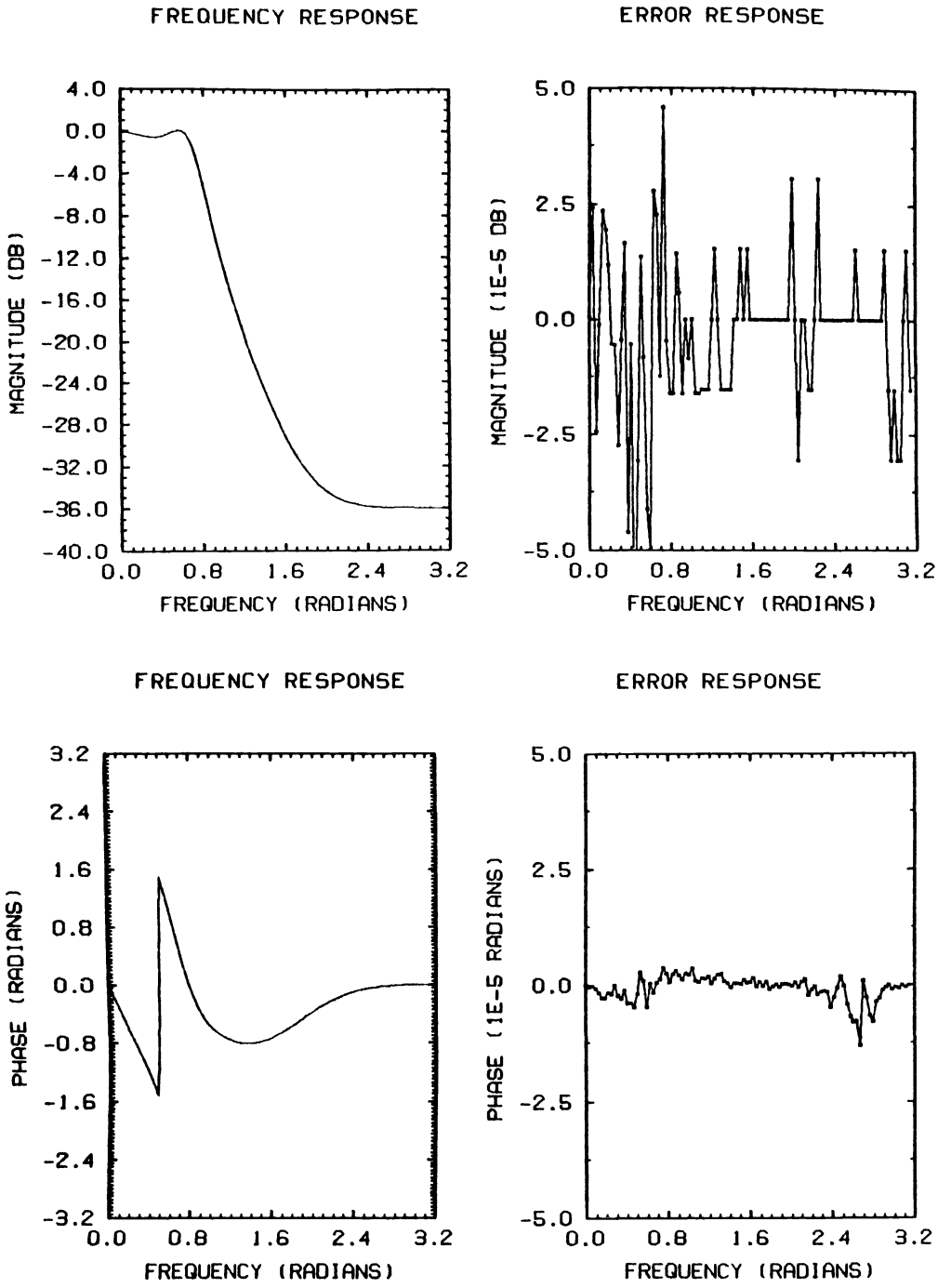


FIG. 10. *Magnitude and phase errors of the reduced sensitivity system.*

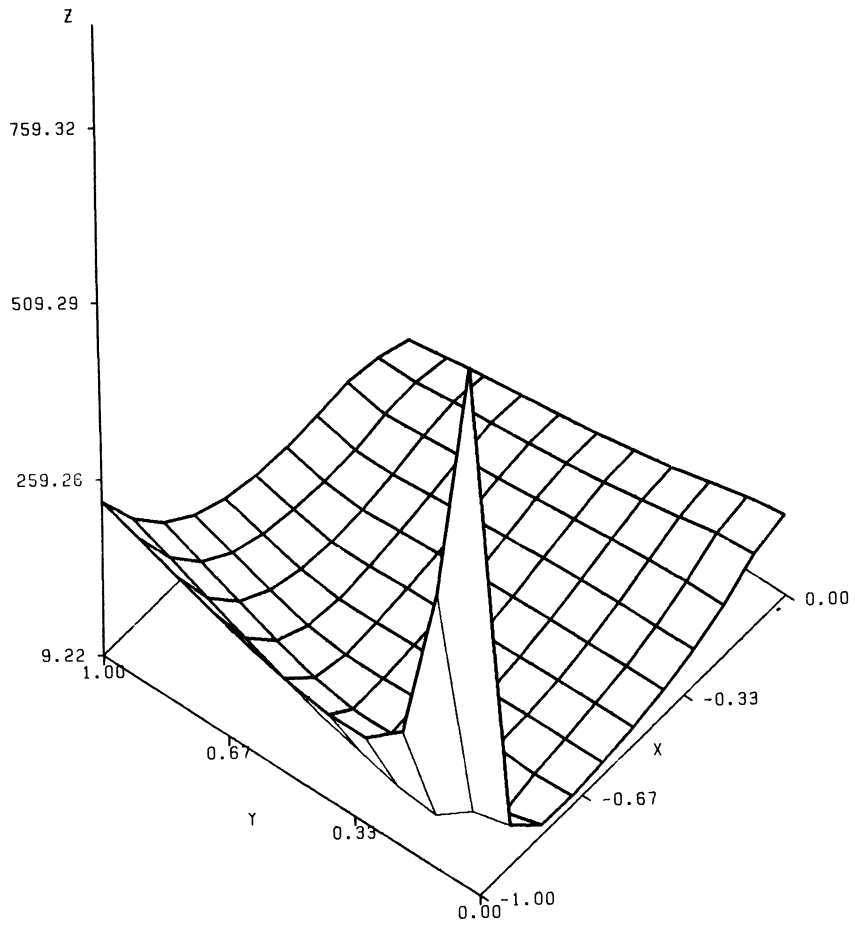


FIG. 11. *Sensitivity of fifth-order implementations.*

not approach sensitivity of the optimal form. For verification, consider the example used by Kawamata and Higuchi. This example has transfer function

$$H(z) = d + \frac{N(z)}{D(z)} \tag{51}$$

where

$$d = .00000869,$$

$$N(z) = .000627(.108543\,z^{-1} + .0067z^{-2} + .104730z^{-3} + .00193z^{-4}),$$

$$D(z) = 1 - 3.826389z^{-1} + 5.516625z^{-2} - 3.551099z^{-3} + .86102z^{-4}.$$

The transfer function is extremely narrow-band. The optimal form has sensitivity measure S_2, equal to 58.327987, as compared to the direct form II, which has a sensitivity of 18,933,029.42. Clearly, the direct form II would not normally be used when accuracy is important, as it is many times (3.25×10^5) more sensitive than the optimal form. Placing a double pole/zero pair at $-.98$ on the real axis in the z-plane causes a reduction in coefficient sensitivity of the direct form II to 1,857,725.66, an improvement of one order

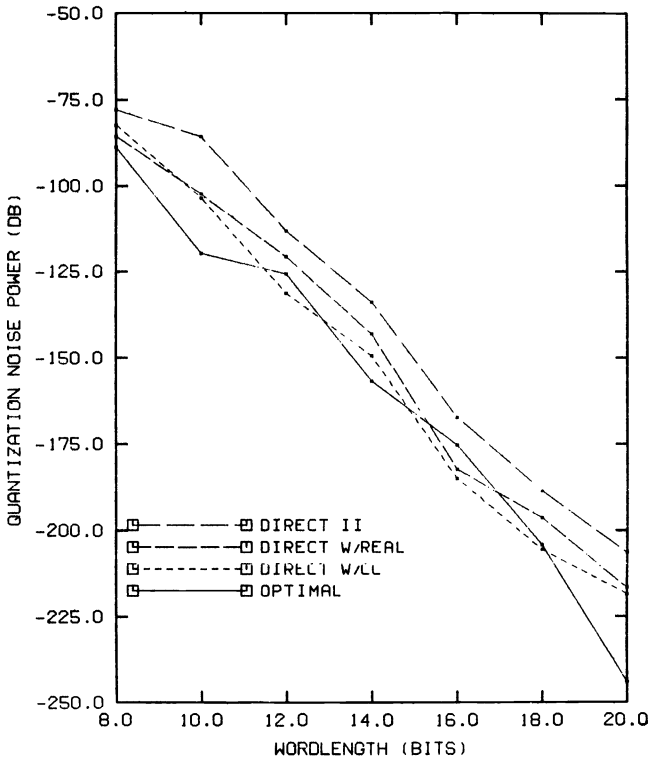


FIG. 12. *Comparison of the optimal and reduced sensitivity forms.*

of magnitude (10.2). However, the reduced sensitivity is still not close to that of the optimal form; the sensitivity is four orders of magnitude greater.

8. Conclusions. Here we have shown the relationship between the sensitivity measure and the output quantization noise power. Efficient methods for calculating both the sensitivity and the output quantization noise power are given, and comparisons to previous computational procedures are made. Computational problems are discussed, and methods to alleviate them are presented. Next, the direct relationship between the pole and zero sensitivities and the sensitivity measure is exploited to reduce the system output quantization noise power of low-pass, direct form II digital filters by the introduction of judiciously placed pole/zero cancellation pair(s). Low-pass, relatively low-order filters designed this way are competitive with optimal designs. These cancellation pair(s) do not affect the system transfer function. Even though our approach also works for narrow band, high-order filters, the corresponding designs are in that case not competitive with the optimal design.

REFERENCES

[1] A. A. BEEX, *Efficient generation of ARMA cross covariance sequences*, IEEE ICASSP '85 Proceedings, March 1985, pp. 327–330.

[2] B. W. BOMAR AND J. C. HUNG, *Minimum roundoff noise digital filters with some power-of-two coefficients*, IEEE Trans. Circuits and Systems, 31 (1984), pp. 833–840.

[3] B. W. BOMAR, *New second-order state-space structures for realizing low roundoff noise digital filters*, IEEE Trans. Acoust., Speech Signal Process., 33 (1985), pp. 106–110.

[4] B. W. BOMAR, *Computationally efficient low roundoff noise second-order state-space structures*, IEEE Trans. Circuits and Systems, 33 (1986), pp. 35–41.

[5] A. G. CONSTANTINIDES AND R. A. VALENZUELA, *A class of efficient interpolators and decimators with applications in transmultiplexers*, Proc. IEEE Internat. Symposium Circuits Syst., Rome, Italy, May 1982, pp. 260–263.

[6] ———, *An efficient and modular transmultiplexer design*, IEEE Trans. Comm., 30 (1982), pp. 1629–1641.

[7] 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.

[8] J. DUGRE, A. A. BEEX, AND L. L. SCHARF, *Generating covariance sequences and the calculation of quantization and rounding error variances in digital filters*, IEEE Trans. Acoust., Speech Signal Process., 28 (1980), pp. 102–104.

[9] A. FETTWEIS, *On sensitivity and roundoff noise in wave digital filters*, IEEE Trans. Acoust., Speech, Signal Process., 22 (1974), pp. 383–384.

[10] ———, *Wave digital lattice filters*, Internat. J. Circuit Theory Appl., 2 (1974), pp. 203–211.

[11] S. Y. HWANG, *Minimum uncorrelated unit noise in state-space digital filtering*, IEEE Trans. Acoust., Speech Signal Process., 25 (1977), pp. 273–281.

[12] L. B. JACKSON, *Roundoff noise bounds derived from coefficient sensitivities for digital filters*, IEEE Trans. Circuits and Systems, 23 (1976), pp. 481–485.

[13] ———, *Digital Filters and Signal Processing*, Kluwer Academic Press, Hingham, MA, 1986.

[14] L. B. JACKSON, A. G. LINDGREN, AND Y. KIM, *Optimal synthesis of second-order state-space structures for digital filters*, IEEE Trans. Circuits and Systems, 26 (1979), pp. 149–153.

[15] J. F. KAISER, *Digital filters*, in System Analysis by Digital Computer, F. F. Kuo and J. F. Kaiser, John Wiley, New York, 1966, Chapter 7.

[16] ———, *Some practical considerations in the realization of linear digital filters*, Proc. 3rd Allerton Conf. Circuit System Theory, October 20–22, 1965, pp. 621–633.

[17] M. KAWAMATA AND T. HIGUCHI, *A unified approach to the optimal synthesis of fixed-point state-space digital filters*, IEEE Trans. Acoust., Speech Signal Process., 33 (1983), pp. 911–920.

[18] J. B. KNOWLES AND E. M. OLCAYTO, *Coefficient accuracy and digital filter response*, IEEE Trans. Circuit Theory, 15 (1968), pp. 31–41.

[19] L. C. LUDEMAN, *Fundamentals of Digital Signal Processing*, Harper and Row, New York, 1986.

[20] J. L. MELSA, *Computer Programs for Computational Assistance in the Study of Linear Control Theory*, McGraw-Hill, New York, 1970, pp. 39–55, 95–97, 119–120.

[21] S. K. MITRA AND R. J. SHERWOOD, *Canonic realizations of digital filters using the continued fraction expansion*, IEEE Trans. Audio Electroacoust., 20 (1972), pp. 185–194.

[22] C. T. MULLIS AND R. A. ROBERTS, *Synthesis of minimum roundoff noise fixed point digital filters*, IEEE Trans. Circuits and Systems, 23 (1976), pp. 551–562.

[23] ———, *Roundoff noise in digital filters: frequency transformations and invariants*, IEEE Trans. Acoust., Speech Signal Process., 24 (1976), pp. 538–550.

[24] A. V. OPPENHEIM AND R. W. SCHAFER, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975, pp. 166–171, 186–187, 214–216, 443, 562–570.

[25] L. R. RABINER AND BERNARD GOLD, *Theory and Applications of Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[26] D. V. B. RAO, *Analysis of coefficient quantization errors in state-space digital filters*, IEEE Trans. Acoust., Speech Signal Process., 34 (1986), pp. 131–139.

[27] R. A. ROBERTS AND C. T. MULLIS, *Digital Signal Processing*, Addison-Wesley, Reading, MA, 1987.

[28] V. TAVSANOGLU AND L. THIELE, *Optimal design of state-space digital filters by simultaneous minimization of sensitivity and roundoff noise*, IEEE Trans. Circuits and Systems, 31 (1984), pp. 884–888.

[29] P. P. VAIDYANATHAN, S. K. MITRA, AND Y. NEUVO, *A new approach to the realization of low-sensitivity IIR digital filters*, IEEE Trans. Acoust., Speech Signal Process., 34 (1986), pp. 350–361.

[30] C. F. VAN LOAN AND G. H. GOLUB, *Matrix Computations*, The John Hopkins University Press, Baltimore, MD, 1983, p. 3.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 1, January 1988

010

ACCURATE SOLUTIONS OF ILL-POSED PROBLEMS IN CONTROL THEORY*

JAMES DEMMEL† AND BO KÅGSTRÖM‡

Abstract. Computable, guaranteed error bounds are presented for controllable subspaces and uncontrollable modes, unobservable subspaces and unobservable modes, supremal (A, C) invariant subspaces in $\ker D$, supremal (A, C) controllability subspaces in $\ker D$, the uncontrollable modes within the supremal (A, C) invariant subspace in $\ker D$, and invariant zeros. In particular the bounds apply in the nongeneric case when the solutions are ill-posed. This is done by showing that all these features are eigenspaces and eigenvalues of certain singular matrix pencils, which means they may all be computed by a single algorithm to which a perturbation theory for general singular matrix pencils can be applied. Numerical examples are included.

Key words. controllability, observability, generalized eigenproblem

AMS(MOS) subject classifications. 93B35, 93B40, 65H15

1. Introduction.

We consider the general linear system

$$B\dot{x} = Ax + Cu, \qquad y = Dx + Fu \tag{1}$$

where A and B are n by n, C is n by k, F is p by m, and D is p by n. We assume B is nonsingular for reasons given in § 4.

Associated with (1) are the following features we would like to compute: uncontrollable subspace, unobservable subspace, maximal (A, C) invariant subspace in $\ker D$, maximal (A, C) controllability subspace in $\ker D$, uncontrollable modes, unobservable modes, invariant zeros, and uncontrollable modes within the maximal (A, C) invariant subspace in $\ker D$. (These features will be defined more precisely in § 4.)

All of these features may be ill-posed, i.e., arbitrarily small changes in A, B, C, D, and F may change them completely. If, for example, the system has r uncontrollable modes, almost any perturbation of A or C will make them disappear. However, if we restrict the perturbed system to have the same structure as the unperturbed one (e.g., to have r uncontrollable modes), then they will vary continuously with A, B, C, D, and F. The set of systems (A, B, C, D, F) with a fixed structure forms a lower-dimensional surface in the space of all systems.

Despite this potential ill-posedness these features are important in practice because the physical structure of a system may force it to have a fixed structure, in which case we would like to compute it accurately. It is also of interest to know if a system is close to one with a given structure, as the system may display an instability associated with that structure. For example, if a system is close to one with an uncontrollable mode, it may take very large feedback to move that mode.

* Received by the editors April 13, 1987; accepted for publication October 1, 1987. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986. A preliminary version of this paper appeared in *Proc. 25th IEEE Conference on Decision and Control* (© 1986 IEEE), December 10–12, 1986, Athens, Greece.

† Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. The work of this author was supported by the National Science Foundation under grant 8501708 and a Presidential Young Investigator Award.

‡ Institute for Information Processing, University of Umeå, S-901 87 Umeå, Sweden. The work of this author was supported by Swedish Natural Science Research Council contract NFR-S-FU1840-101, and Swedish National Board for Technical Development contract STU-84-5481.

There are a number of good algorithms available to compute the features listed above [8], [9], [15], [16]. They share two important properties. First, they are backwards stable, i.e., they compute the features exactly for a slight perturbation of the system given as input. The user may limit the size of the perturbation the algorithm will permit. Second, they attempt to find a slightly perturbed system with as much structure as possible within the user's size limit on the perturbation. By "as much structure as possible" we mean the most uncontrollable modes, the largest unobservable subspace, etc. Thus, for example, if the system (A, C) is sufficiently close to one (A', C') with three uncontrollable modes, but rather farther from a system with more, the algorithm will compute the controllable subspace of dimension $n - 3$ and three uncontrollable modes of (A', C'). Another way to say it is that the algorithm projects the system (A, C) onto a nearby system (A', C') on the surface of systems with three uncontrollable modes, which it analyzes exactly.

Thus, these algorithms are appropriate in situations where either a system is supposed to have a fixed structure, or the user is interested in knowing if the system is close to one with a fixed structure.

In this paper we analyze the accuracy of these algorithms. In light of the preceding discussion, the perturbation theory we develop answers the following question: If two systems have the same structure, how does the distance between their computed features depend on the distance between the systems? For example, if two systems (A, C) and (A', C') both have three uncontrollable modes, how does the angle between their controllable subspaces depend on the distance from A to A' and C to C'?

Our approach is as follows. First we show that all the algorithms for the features mentioned above are special cases of a single algorithm for computing the Kronecker structure of a matrix pencil $H - \lambda G$. All the features that are subspaces are reducing subspaces (or projections of reducing subspaces) and all the modes and zeros are generalized eigenvalues of particular pencils $H - \lambda G$ whose entries depend on A, B, C, D, and F. Reducing subspaces are the natural generalizations of invariant subspaces for the standard eigenproblem $H - \lambda I$ to the generalized eigenproblem $H - \lambda G$ [17]. For a more complete account of this material, see [18].

Second, we apply a perturbation theory for reducing subspaces and generalized eigenvalues of arbitrary pencils [4] to the particular pencils of the last paragraph. This perturbation theory supplies the following information: if two systems (A, B, C, D, F) and (A', B', C', D', F') have the same structure, and if the distance d between them is less than an upper bound Δ which is computable straightforwardly from the entries of (A, B, C, D, F), then the distance between their features is less than $\kappa \cdot d$, where κ is also straightforward to compute. (We will discuss Δ, κ, and the distance measures we use in a later section.) In other words, this theory provides guaranteed computable upper bounds on the error in computed features.

We have implemented an improved version of algorithm RGQZD [5] (a unitary version of the RGSVD algorithm [9]) for computing the reducing subspaces and generalized eigenvalues of a matrix pencil $H - \lambda G$, as well as computing the quantities Δ and κ in the perturbation theorem. Preliminary numerical experiments are in agreement with the perturbation theory and also show that the bounds are realistic. We report on these results in § 6.

The paradigm used in this paper, projecting an ill-conditioned problem onto a surface of problems with a fixed structure to improve the conditioning, is a regularization technique common in numerical analysis. The canonical example is using the pseudoinverse to solve nearly rank deficient least squares problems: setting the small singular values to zero improves the conditioning by projecting the matrix onto the surface of rank deficient ones.

The rest of the paper is organized as follows. Section 2 defines notation and explains the distance measures we use later. Section 3 records some standard facts about the Kronecker Canonical Form and reducing subspaces. Section 4 shows how all the control problems in the Introduction may be expressed as reducing subspaces or generalized eigenvalues of matrix pencils. Section 5 explains the perturbation theory for reducing subspaces and generalized eigenvalues. Section 6 contains numerical examples.

2. Notation. $\|x\|$ will denote the Euclidean norm of the vector x. $\|A\|$ will denote the matrix norm induced by the Euclidean vector norm. $\|A\|_E$ will denote the Frobenius norm, and $\|(A, B)\|_E^2 = \|A\|_E^2 + \|B\|_E^2$. $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ $(=\|A\|)$ will denote the smallest and largest singular values of the matrix A, respectively. $\kappa(A)$ will denote the condition number $\sigma_{\max}(A)/\sigma_{\min}(A)$ of the matrix A; this applies to nonsquare A as well. $A \otimes B$ will denote the Kronecker product of the two matrices A and B: $A \otimes B = [A_{ij} \cdot B]$. Rows (A), columns (A), and rank (A) will denote the number of rows, number of columns, and rank of A, respectively. Let col$A$ denote the column vector formed by taking the columns of A and stacking them atop one another from left to right. Thus if A is m by n, col$A$ is mn by 1 with its first m entries being column 1 of A, its second m entries being column 2 of A, and so on. $\mathbf{R}(X)$ is the space spanned by the columns of X and ker X is the null space of X. The (largest) angle between two subspaces $\mathbf{X}_1$ and $\mathbf{X}_2$ is given by

$$\theta_{\max}(\mathbf{X}_1, \mathbf{X}_2) = \max_{x_1 \in \mathbf{X}_1} \min_{x_2 \in \mathbf{X}_2} \theta(x_1, x_2)$$

where $\theta(x_1, x_2)$ is the acute angle between the nonzero vectors x_1 and x_2.

3. The Kronecker Canonical Form. In this section we briefly review the Kronecker Canonical Form (KCF), reducing subspaces, and an upper triangular canonical form with the same information as the KCF but which may be computed stably. The KCF is a generalization of the Jordan Canonical Form for the standard eigenproblem $H - \lambda I$ to the generalized eigenproblem $H - \lambda G$. Like the Jordan form, the KCF cannot be computed stably so instead we compute an upper triangular canonical form we call GUPTRI (for generalized upper triangular) form which generalizes the Schur canonical form for the standard eigenproblem. Reducing subspaces generalizes the notion of invariant subspaces.

The KCF is defined as follows. Let H and G be m-by-n matrices. Then there exist nonsingular matrices P and Q such that

$$P^{-1}(H - \lambda G)Q = S - \lambda T \tag{2}$$

is block diagonal: $S = \text{diag}(S_{11}, \cdots, S_{bb})$ and $T = \text{diag}(T_{11}, \cdots, T_{bb})$. We can group the columns of P into blocks corresponding to the blocks of $S - \lambda T$: $P = [P_1 | \cdots | P_b]$ where P_i is m by m_i, m_i being the number of rows of $S_{ii} - \lambda T_{ii}$. Similarly we can write $Q = [Q_1 | \cdots | Q_b]$ where Q_i is n by n_i, n_i being the number of columns of $S_{ii} - \lambda T_{ii}$. Each block $S_{ii} - \lambda T_{ii}$ must be of one of the following four forms:

$$J_j(\lambda_0) \equiv \begin{bmatrix} \lambda_0 - \lambda & 1 & & \\ & \ddots & \ddots & \\ & & & 1 \\ & & & \lambda_0 - \lambda \end{bmatrix} \quad \text{or} \quad N_j \equiv \begin{bmatrix} 1 & -\lambda & & \\ & \ddots & \ddots & \\ & & & -\lambda \\ & & & 1 \end{bmatrix}.$$

$J_j(\lambda_0)$ is simply a j-by-j Jordan block. λ_0 is called a finite eigenvalue of the pencil. The j-by-j block N_j corresponds to an infinite eigenvalue of multiplicity equal to the dimension

of the block. The blocks of finite and infinite eigenvalues together constitute the *regular* part of the pencil:

$$L_j \equiv \begin{bmatrix} -\lambda & 1 & & \\ & \ddots & \ddots & \\ & & -\lambda & 1 \end{bmatrix} \quad \text{or} \quad L_j^T \equiv \begin{bmatrix} -\lambda & & \\ 1 & \ddots & \\ & \ddots & -\lambda \\ & & 1 \end{bmatrix}.$$

The j by $j+1$ block L_j is called a singular block of minimal right (or column) index j. It has a one-dimensional right null space for any λ. The $j+1$ by j block L_j^T is called a singular block of minimal left (or row) index j. It has a one-dimensional left null space for any λ. The left and right singular blocks together constitute the *singular* part of the pencil.

If a pencil only has a regular part in its KCF, it is called *regular*. $H - \lambda G$ is regular if and only if it is square and its determinant $\det(H - \lambda G)$ is not identically zero. Otherwise, there is at least one singular block L_j or L_j^T in the KCF of $H - \lambda G$ and it is called *singular*.

In the regular case, $H - \lambda G$ has n generalized eigenvalues which may be finite or infinite. The diagonal blocks of $S - \lambda T$ partition the spectrum of $H - \lambda G$ as follows:

$$\sigma \equiv \sigma(H - \lambda G) = \bigcup_{i=1}^{b} \sigma(S_{ii} - \lambda T_{ii}) \equiv \bigcup_{i=1}^{b} \sigma_i.$$

The subspaces $\mathbf{P}_i$ and $\mathbf{Q}_i$ spanned by P_i and Q_i are called *left and right deflating subspaces* of $H - \lambda G$ corresponding to the part of the spectrum σ_i [11], [17]. As shown in [12], a pair of subspaces $\mathbf{P}$ and $\mathbf{Q}$ is deflating for $H - \lambda G$ if $\mathbf{P} = H\mathbf{Q} + G\mathbf{Q}$ and $\dim(\mathbf{Q}) = \dim(\mathbf{P})$. Deflating subspaces are determined uniquely by the partitioning $\sigma = \cup_{i=1}^{b} \sigma_i$. Different choices of the P_i and Q_i will span the same spaces $\mathbf{P}_i$ and $\mathbf{Q}_i$.

The situation is not as simple in the singular use. The following example shows that the spaces $\mathbf{P}_i$ and $\mathbf{Q}_i$ spanned by block diagonalizing P_i and Q_i may no longer all be well defined:

$$(3) \quad P(S - \lambda T)Q^{-1} = \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} -\lambda & 1 & 0 \\ 0 & 0 & 1-\lambda \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & x \\ 0 & 1 & x \\ 0 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} -\lambda & 1 & 0 \\ 0 & 0 & 1-\lambda \end{bmatrix}.$$

As x grows large, the space spanned by Q_2 (the last column of Q) can become arbitrarily close to the space spanned by Q_1 (the first two columns of Q). Similarly the space spanned by P_2 (the last column of P) can become arbitrarily close to the space spanned by P_1 (the first column of P). Thus we must modify the notion of deflating subspace used in the regular case, since these subspaces are no longer all well defined.

The correct concept to use is *reducing subspace*, as introduced in [17]. $\mathbf{P}$ and $\mathbf{Q}$ are left and right reducing subspaces for $H - \lambda G$ if $\mathbf{P} = H\mathbf{Q} + G\mathbf{Q}$ and $\dim(\mathbf{P}) = \dim(\mathbf{Q}) - \#(L_j \text{ blocks in the KCF of } H - \lambda G)$. In terms of the KCF, $\mathbf{Q}$ is spanned by all the Q_i where $S_{ii} - \lambda T_{ii} = L_j$ plus the Q_i for any subset $\sigma_1 \subseteq \sigma$ of the regular part. $\mathbf{P}$ is spanned by the corresponding P_i. Thus there is a pair of reducing subspaces for every subset (including empty and full) of the spectrum of the regular part of $H - \lambda G$. When the subset of the spectrum is empty, we will call the corresponding pair of reducing subspaces *minimal*, and when the subset consists of all the eigenvalues, we will call the reducing subspaces *maximal*. In example (3), P_1 and Q_1 span the minimal reducing subspaces.

Just as the Jordan form cannot be computed stably and is the wrong way to compute invariant subspaces, the KCF is the wrong way to compute reducing subspaces. Instead,

to compute invariant subspaces we limit ourselves to unitary transformations and the Schur canonical form. There is an analogous generalized upper triangular form (which we call GUPTRI) for matrix pencils [17]: there exist unitary P and Q such that

$$P^{-1}(H-\lambda G)Q=\begin{bmatrix} H_r-\lambda G_r & * & * \\ 0 & H_{\text{reg}}-\lambda G_{\text{reg}} & * \\ 0 & 0 & H_l-\lambda G_l \end{bmatrix}. \tag{4}$$

Here $H_r - \lambda G_r$ has only L_j blocks in its KCF, $H_{\text{reg}} - \lambda G_{\text{reg}}$ is regular, and $H_l - \lambda G_l$ has only L_j^T blocks in its KCF. The KCF of $H - \lambda G$ has the same blocks as in $H_r - \lambda G_r$, $H_{\text{reg}} - \lambda G_{\text{reg}}$, and $H_l - \lambda G_l$. (The order of the blocks on the diagonal is essential for this statement to be true.) Furthermore, $H_{\text{reg}} - \lambda G_{\text{reg}}$ may be chosen upper triangular with its eigenvalues appearing in any order. It is easy to determine the reducing subspaces from GUPTRI. Suppose $H_r - \lambda G_r$ is m_i by n_i and that the eigenvalues σ_1 that correspond to the desired reducing subspace appear in the upper left i-by-i corner of $H_{\text{reg}} - \lambda G_{\text{reg}}$ (i may be 0). Then the left and right reducing subspaces corresponding to σ_1 are spanned by the first $m_i + i$ columns of P and first $n_i + i$ columns of Q, respectively. In this case we write the decomposition (4) as

$$P^{-1}(H-\lambda G)Q=\begin{bmatrix} H_{11}-\lambda G_{11} & H_{12}-\lambda G_{12} \\ 0 & H_{22}-\lambda G_{22} \end{bmatrix} \tag{5}$$

where $H_{11} - \lambda G_{11}$ is $m_1 + i$ by $n_1 + i$.

A number of workers have developed stable algorithms for computing GUPTRI (or similar forms) [8], [9], [15], [17], [20]. In addition, more efficient algorithms have been developed when $H - \lambda G$ takes on special structures pertinent to control theory [6], [16].

4. Reducing subspaces in control theory. In this section we show that all the subspaces listed in the Introduction are reducing subspaces of particular matrix pencils (or projections onto certain components of reducing subspaces), and that all the modes and zeros are generalized eigenvalues of those pencils. The point of view is originally due to Van Dooren [17], [18]. For a more thorough discussion of these control problems see [21].

First we consider the controllable subspace and uncontrollable modes of the system

$$B\dot{x}=Ax+Cu. \tag{6}$$

We assume the pencil $A - \lambda B$ is regular and B nonsingular in particular. Following [16] we define the following:

DEFINITION 1. The *controllable subspace* of (6) is the right deflating subspace $\mathbf{Q}$ of the smallest pair of deflating subspaces $\mathbf{P}$ and $\mathbf{Q}$ of $A - \lambda B$ satisfying $\mathbf{R}(C) \subseteq \mathbf{P}$. The *uncontrollable modes* of (6) are the eigenvalues of $A - \lambda B$ corresponding to the complementary deflating subspace.

When B is the identity matrix, this definition reduces to the usual one (in particular $\mathbf{P} = \mathbf{Q}$). The definition makes sense when B is singular and $A - \lambda B$ regular, but one important property of complete controllability is lost in this case: pole assignability. The feedback $u = Kx$ leads to the pencil $A + CK - \lambda B$, where B is still singular. If this new pencil is regular, it must have as many N_j blocks of infinite eigenvalues in its KCF as the original pencil. Worse, feedback may lead to a singular pencil with nonphysical or no solutions of the corresponding differential equation [7]. For example, if

$$A-\lambda B=\begin{bmatrix} 1 & \\ & -\lambda \end{bmatrix} \quad \text{and} \quad C=\begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad K=[-1,0],$$

then the new pencil

$$A + CK - \lambda B = \begin{bmatrix} 0 & \\ & -\lambda \end{bmatrix}$$

is singular. However, as long as B is nonsingular pole assignability and controllability are equivalent as before.

An extension of the definitions of controllability and observability to the case of singular B is given in [2]. These definitions also permit interpretations in terms of the KCF. For example, the "nonsingular" part of the system is controllable if and only if there are no finite eigenvalues. We intend to extend our results to the case of singular B using these definitions in a future paper.

We may now state the following.

THEOREM 1. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be the minimal left and right reducing subspaces of*

$$[C|A - \lambda B], \tag{7}$$

where B is invertible. Then

(i) $\mathbf{Q} = \mathbf{R}([\begin{smallmatrix} I_k \\ 0 \end{smallmatrix}]) + \mathbf{R}([\begin{smallmatrix} 0 \\ Q \end{smallmatrix}])$ *and* $\mathbf{R}(Q)$ *is the controllable subspace. In other words, the bottom n components of any basis of* $\mathbf{Q}$ *span the controllable subspace. The controllable subspace also equals* $B^{-1}\mathbf{P}$.

(ii) *The eigenvalues of the regular part of* $[C|A - \lambda B]$ *are the uncontrollable modes.*

Proof. We show first that if $\mathbf{P}$ and $\mathbf{Q}$ are any pair of reducing subspaces, then $\mathbf{P}$ and the last n components of $\mathbf{Q}$ are a pair of left and right deflating subspaces of $A - \lambda B$ with $R(C) \subseteq \mathbf{P}$. Since $A - \lambda B$ is regular, it as well as $[C|A - \lambda B]$ has full row rank for almost all λ and so (7) cannot have any L_j^T blocks in its KCF. Therefore the number of L_j blocks in the KCF must be columns $(C) = k = \dim(\mathbf{Q}) - \dim(\mathbf{P})$. Let $\mathbf{P} = \mathbf{R}(P)$, where P is n by c and $\mathbf{Q} = R([\begin{smallmatrix} Q_1 \\ Q_2 \end{smallmatrix}])$ where Q_1 is k by $k + c$ and Q_2 is n by $k + c$. The definition of reducing subspace implies

$$\mathbf{P} = [C|A]\mathbf{Q} + [0|B]\mathbf{Q} = \mathbf{R}(CQ_1 + AQ_2) + \mathbf{R}(BQ_2). \tag{8}$$

The rank of Q_2 must equal $\dim(\mathbf{P})$, since if it were less $[\begin{smallmatrix} Q_1 \\ Q_2 \end{smallmatrix}]$ would not have full column rank, and if it were more $\dim(\mathbf{R}(BQ_2))$ would be larger than $\dim(\mathbf{P})$. Therefore we may assume

$$\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = \begin{bmatrix} I_k & 0 \\ 0 & Q_{22} \end{bmatrix}$$

where Q_{22} is n by c and of rank c. Thus (8) becomes

$$\mathbf{P} = \mathbf{R}(C) + \mathbf{R}(AQ_{22}) + \mathbf{R}(BQ_{22})$$

implying that $\mathbf{R}(C) \subseteq \mathbf{P}$ and that $\mathbf{P}$ and $\mathbf{R}(Q_{22})$ form a pair of deflating subspaces of $A - \lambda B$.

Conversely, it is easy to see that if $\mathbf{R}(L)$ and $\mathbf{R}(R)$ are a pair of deflating subspaces of $A - \lambda B$ with $\mathbf{R}(C) \subseteq \mathbf{R}(L)$, then $\mathbf{R}(L)$ and $\mathbf{R}([\begin{smallmatrix} I & 0 \\ 0 & R \end{smallmatrix}])$ are a pair of reducing subspaces of (7). Thus there is a one-to-one correspondence between reducing subspaces of (7) and deflating subspaces of $A - \lambda B$ where the left deflating subspace contains $\mathbf{R}(C)$. In particular, the minimal reducing subspace corresponds to the smallest deflating subspace in Definition 1. It is easy to see $B^{-1}\mathbf{R}(L) = \mathbf{R}(R)$.

It remains to prove that the eigenvalues of the regular part of (7) are the uncontrollable modes. Let $\mathbf{R}(L)$ and $\mathbf{R}([\begin{smallmatrix} I & 0 \\ 0 & R \end{smallmatrix}])$ be the minimal left and right reducing subspaces. Choose L_1 and R_1 so that

$$P = [L|L_1] \quad \text{and} \quad Q = \begin{bmatrix} I & 0 & 0 \\ 0 & R & R_1 \end{bmatrix}$$

are nonsingular (in fact they may be chosen unitary). Then

$$P^{-1}[C|A - \lambda B]Q = \begin{bmatrix} C_1 & A_{11} - \lambda B_{11} & A_{12} - \lambda B_{12} \\ 0 & 0 & A_{22} - \lambda B_{22} \end{bmatrix}.$$

This is in GUPTRI form. From the previous discussion, the KCF of $[C_1|A_{11} - \lambda B_{11}]$ consists only of L_j blocks (i.e., it is completely controllable), and $A_{22} - \lambda B_{22}$ is regular, and its spectrum consists of the uncontrollable modes. □

Because of the structure of the minimal right reducing subspace of (7), the largest angle between two such subspaces will be the largest angle between the two spaces spanned by their bottom n components. We use this fact to convert our bound on the angle between perturbed reducing subspaces to a bound for perturbed controllable subspaces.

Next we consider the unobservable subspace and modes. The system we consider is

$$\begin{aligned} B\dot{x} &= Ax, \\ y &= Dx. \end{aligned} \tag{9}$$

Following [16] again we define the following.

DEFINITION 2. The *unobservable subspace* of (9) is the right deflating subspace $\mathbf{Q}$ of the largest pair of deflating subspaces $\mathbf{P}$ and $\mathbf{Q}$ of $A - \lambda B$ satisfying $\mathbf{Q} \subseteq \ker D$. The *unobservable modes* are the eigenvalues of $A - \lambda B$ corresponding to $\mathbf{P}$ and $\mathbf{Q}$.

We may now state the following.

THEOREM 2. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be the maximal left and right reducing subspaces of*

$$\begin{bmatrix} A - \lambda B \\ D \end{bmatrix} \tag{10}$$

where B *is invertible. Then*

(i) $\mathbf{Q}$ *is the unobservable subspace. If* $\mathbf{P} = \mathbf{R}([\begin{smallmatrix} P_1 \\ P_2 \end{smallmatrix}])$ *where* P_1 *has* n *rows, then* $P_2 = 0$ *and* $B^{-1}\mathbf{R}(P_1)$ *is also the unobservable subspace.*

(ii) *The unobservable modes are the eigenvalues of the regular part of* (10).

Proof. The proof will use duality. Changing A to $B^{-1}A$ and B to the identity changes neither the unobservable space or unobservable modes of (9), nor the right reducing subspaces or eigenvalues of (10). Therefore assume without loss of generality that $B = I$. In this case by duality

$$\begin{aligned} &\text{unobservable subspace of } C, A \\ &\quad = (\text{controllable subspace of } A^T, C^T)^{\perp} \\ &\quad = (\text{minimal left reducing subspace of } [C^T|A^T - \lambda I])^{\perp} \\ &\quad = (\text{minimal left reducing subspace of } [A^T - \lambda I|C^T])^{\perp} \\ &\quad = \text{maximal right reducing subspace of } [\begin{smallmatrix} A - \lambda I \\ C \end{smallmatrix}]. \end{aligned}$$

(The last equality follows from the fact that transposing a pencil exchanges L_j and L_j^T blocks without changing the other blocks in the KCF.) The relationship between unobservable modes and eigenvalues also follows from duality. □

Now we turn to (A, C) invariant and controllability subspaces in ker D and their generalizations. Consider the system

$$B\dot{x} = Ax + Cu, \tag{11}$$
$$y = Dx.$$

DEFINITION 3. $\mathbf{Q}$ is an (A, B, C) invariant subspace in ker D if there is another subspace $\mathbf{P}$ of the same dimensions as $\mathbf{Q}$ satisfying

$$A\mathbf{Q} \subseteq \mathbf{P} + \mathbf{R}(C),$$
$$B\mathbf{Q} \subseteq \mathbf{P},$$
$$\mathbf{Q} \subseteq \ker D.$$

It is easy to see from the definition that if $\mathbf{Q}$ and $\mathbf{P}$ satisfy it, there is a feedback matrix K such that $\mathbf{P}$ and $\mathbf{Q}$ form a pair of deflating subspaces for the pencil $A + CK - \lambda B$. It is also easy to see that if $B = I$, this definition reduces to the usual one for (A, C) invariant subspaces in ker D. We now need to establish the following.

PROPOSITION 1. *Suppose B is invertible. Then there is a supremal (A, B, C) invariant subspace in* ker D *containing all others.*

Proof. We reduce to the case where $B = I$. $\mathbf{Q}$ is an (A, B, C) invariant subspace in ker D with other subspace $\mathbf{P}$ if and only if it is a $(B^{-1}A, I, B^{-1}C)$ invariant subspace in ker D with other subspace $B^{-1}\mathbf{P} = \mathbf{Q}$. Since this new problem has a supremal subspace so does the original one. □

We will call the supremal (A, B, C) invariant subspace in ker $D\mathbf{V}^*$ if A, B, C and D are clear from context.

DEFINITION 4. $\mathbf{Q}$ is an (A, B, C) controllability subspace in ker D if it is an (A, B, C) invariant subspace in ker D and for any set of dim $(\mathbf{Q})$ scalars $\{\lambda_i\}$ a feedback matrix K can be chosen so that $\mathbf{Q}$ is a right deflating subspace of $A + CK - \lambda B$ whose corresponding eigenvalues are $\{\lambda_i\}$.

Note that since this definition refers explicitly to pole assignability, the nonsingularity of B is important. Analogous to Proposition 1, we need to establish the following.

PROPOSITION 2. *Suppose B is invertible. Then there is a supremal (A, B, C) controllability subspace in* ker D *containing all others.*

Proof. As before, we reduce to the case $B = I$. $\mathbf{Q}$ is an (A, B, C) controllability subspace in ker D with other subspace $\mathbf{P}$ if and only if it is a $(B^{-1}A, I, B^{-1}C)$ controllability subspace in ker D with other subspace $B^{-1}\mathbf{P} = \mathbf{Q}$. Since this new problem has a supremal subspace so does the original one. □

We will call the supremal (A, B, C) controllability subspace in ker $D\mathbf{R}^*$ if A, B, C and D are clear from context. To analyze $\mathbf{V}^*$ and $\mathbf{R}^*$ we need the following.

LEMMA 1. *Suppose B is invertible. Let $\mathbf{V}^*$ be the supremal (A, B, C) invariant subspace in* ker D. *Then there exist n by n unitary matrices $Q = [Q_1|Q_2|Q_3]$ and $P = [P_1|P_2|P_3]$, where Q_i and P_i both have n_i columns, such that the decomposition*

$$\begin{bmatrix} P^* & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} C & A - \lambda B \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & Q \end{bmatrix} = \begin{bmatrix} C_1 & A_{11} - \lambda B_{11} & A_{12} - \lambda B_{12} & A_{13} - \lambda B_{13} \\ C_2 & A_{21} & A_{22} - \lambda B_{22} & A_{23} - \lambda B_{23} \\ 0 & 0 & A_{32} - \lambda B_{32} & A_{33} - \lambda B_{33} \\ 0 & 0 & D_2 & D_3 \end{bmatrix} \tag{12}$$

(where A_{ij} is n_i by n_j) has the following properties:

(i) $\mathbf{V}^* = \mathbf{R}(Q_1)$.

(ii) *C_2 has full row rank.*

(iii) $\begin{bmatrix} A_{32} - \lambda B_{32} & A_{33} - \lambda B_{33} \\ D_2 & D_3 \end{bmatrix}$ *has full column rank for all finite* λ.

(iv) *The system with*

$$A' = \begin{bmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{bmatrix}, \qquad B' = \begin{bmatrix} B_{22} & B_{23} \\ B_{32} & B_{33} \end{bmatrix}, \qquad D' = [D_2 D_3]$$

is completely observable.

Proof. Choose P_i and Q_i so that $\mathbf{R}(Q_1) = \mathbf{V}^*$, $\mathbf{R}(P_i)$ is the left deflating subspace for $\mathbf{V}^*$ in Definition 3, $\mathbf{R}(P_2) + \mathbf{R}(P_1) \supseteq \mathbf{R}(C)$ and dim (P_2) is minimal. Thus in

$$P^*C = \begin{bmatrix} P_1^* C \\ P_2^* C \\ P_3^* C \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ 0 \end{bmatrix}$$

it is clear that $P_3^* C = 0$ and C_2 has full row rank. From Definition 3, it is also clear that $A_{31} = P_3^* A Q_1 = 0$, $B_{32} = P_2^* B Q_1 = 0$ and $B_{31} = P_3^* B Q_1 = 0$. Finally $D_1 = D Q_1 = 0$. This explains the structure of (12) and proves claims (i) and (ii).

We prove claim (iii) by contradiction. Since C_2 has full row rank we can choose a feedback matrix $K = [K_1 | K_2 | K_3]$ so that the middle n_2 rows of $A + CK$ are $[0 | X_1 | X_2]$ where X_1 and X_2 can be chosen arbitrarily. Thus if the matrix in claim (ii) did not have full column rank for some λ', we could find matrices X_1 and X_2 such that

$$\begin{bmatrix} X_1 - \lambda' B_{22} & X_2 - \lambda' B_{23} \\ A_{32} - \lambda' B_{32} & A_{33} - \lambda' B_{33} \\ D_2 & D_3 \end{bmatrix}$$

had dependent columns as well. Thus, from Theorem 2 the system with

$$A'' = \begin{bmatrix} X_1 & X_2 \\ A_{32} & A_{33} \end{bmatrix}, \qquad B'' = \begin{bmatrix} B_{22} & B_{23} \\ B_{32} & B_{33} \end{bmatrix}, \qquad D'' = [D_2 | D_3]$$

would have an unobservable subspace. Thus there would be unitary T_1 and T_2 such that

$$T_1 A'' T_2 = \begin{bmatrix} a''_{11} & A''_{12} \\ 0 & A''_{22} \end{bmatrix}, \qquad T_1 B'' T_2 = \begin{bmatrix} b''_{11} & B''_{12} \\ 0 & B''_{22} \end{bmatrix}, \qquad D'' T_2 = [0 | D''_2].$$

This implies that there is an $(A + CK, B, C)$ invariant subspace in ker D of dimension larger than $\mathbf{V}^*$, which contradicts supremality of $\mathbf{V}^*$. Claim (iv) follows immediately from claim (iii). It is easy to see that the proof still goes through if $\mathbf{V}^* = \{0\}$ and $n_i = 0$. □

Now we can prove the following.

THEOREM 3. *Suppose B is invertible. Consider the pencil*

$$\begin{bmatrix} C & A - \lambda B \\ 0 & D \end{bmatrix}. \tag{13}$$

(i) *Let $\mathbf{P}_i$ and $\mathbf{Q}_i$ be the largest left and right reducing subspaces of* (13) *excluding any infinite eigenvalues. Suppose*

$$\mathbf{Q}_i = \mathbf{R}\left(\begin{bmatrix} Q_{i1} \\ Q_{i2} \end{bmatrix}\right) \quad \text{and} \quad \mathbf{P}_i = \mathbf{R}\left(\begin{bmatrix} P_{i1} \\ P_{i2} \end{bmatrix}\right)$$

where Q_{i2} and P_{i1} both have n rows. Then $P_{i2} = 0$ and the supremal (A, B, C) invariant subspace in ker D *is*

$$\mathbf{V}^* = \mathbf{R}(Q_{i2}) = B^{-1}\mathbf{R}(P_{i1})$$

with dim $(\mathbf{V}^*)$ = dim $(\mathbf{P}_i)$.

(ii) *Let P_c and Q_c be the minimal left and right reducing subspaces of* (13). *Suppose*

$$Q_c = \mathbf{R}\left(\begin{bmatrix} Q_{c1} \\ Q_{c2} \end{bmatrix}\right) \quad \textit{and} \quad \mathbf{P}_c = \mathbf{R}\left(\begin{bmatrix} P_{c1} \\ P_{c2} \end{bmatrix}\right)$$

where Q_{c2} and P_{c1} both have n rows. Then $P_{c2} = 0$ and the supremal (A, B, C) controllability subspace in ker D *is*

$$\mathbf{R}^* = \mathbf{R}(Q_{c2}) = B^{-1}\mathbf{R}(P_{c1})$$

with dim $(\mathbf{R}^*)$ = dim (P_c).

(iii) *The finite eigenvalues of* (13) *are the uncontrollable modes in* $\mathbf{V}^*$. *In other words, if any feedback matrix K is chosen so that $A + CK - \lambda B$ has $\mathbf{V}^*$ as a right deflating subspace, it will have as corresponding eigenvalues the finite eigenvalues of* (13).

Proof. There are two cases, $\mathbf{V}^* = \{0\}$ and $\mathbf{V}^* \neq \{0\}$. In the first case the decomposition (12) reduces to

$$\begin{bmatrix} P^* & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} C & A-\lambda B \\ 0 & D \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & Q \end{bmatrix} = \begin{bmatrix} C_2 & A_{22}-\lambda B_{22} & A_{23}-\lambda B_{23} \\ 0 & A_{32}-\lambda B_{32} & A_{33}-\lambda B_{33} \\ 0 & D_2 & D_3 \end{bmatrix}$$

where

$$\begin{bmatrix} A_{32}-\lambda B_{32} & A_{33}-\lambda B_{33} \\ D_2 & D_3 \end{bmatrix}$$

has full column rank for all finite λ. Therefore it can have only infinite eigenvalues and L_j^T blocks in its KCF. Therefore (12) itself will have only L_0 blocks, infinite eigenvalues, and L_j^T blocks in its KCF. Thus $\mathbf{P}_i = \{0\}$ and $Q_{i2} = 0$ as well.

Clearly $\mathbf{R}^* = \{0\}$ in this case as well, and there are no finite eigenvalues.

Now consider $\mathbf{V}^* \neq \{0\}$. We take decomposition (12) and perform the following three transformations on it:

(1) Since C_2 has full row rank, there is a k-by-n feedback matrix K such that $C_2K = -A_{21}$. Postmultiply both sides of (12) by

$$\begin{bmatrix} I_k & K & 0 \\ 0 & I_{n_1} & 0 \\ 0 & 0 & I_{n_2+n_3} \end{bmatrix}$$

to eliminate the A_{21} entry of A.

(2) Since C_2 has full row rank there is a unitary k-by-k matrix $S = [S_1|S_2]$ such that $C_2S = [0|C_{22}]$, where both S_2 and C_{22} have n_2 columns. Write $C_1S = [C_{11}|C_{12}]$ where C_{12} also has n_2 columns. Postmultiply both sides of (12) by $\left[\begin{smallmatrix} S & 0 \\ 0 & I_n \end{smallmatrix}\right]$.

(3) Exchange columns $k - n_2 + 1$ to k with columns $k + 1$ to $k + n$, on both sides of (12). This is equivalent to postmultiplication with a permutation. At the end of these three transformations (12) has become

$$\begin{bmatrix} P^* & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} C & A-\lambda B \\ 0 & D \end{bmatrix}\begin{bmatrix} S_1 & K & S_2 & 0 & 0 \\ 0 & Q_1 & 0 & Q_2 & Q_3 \end{bmatrix}$$

$$= \begin{bmatrix} C_{11} & A_{11}+C_1K-\lambda B_{11} & C_{12} & A_{12}-\lambda B_{12} & A_{13}-\lambda B_{13} \\ 0 & 0 & C_{22} & A_{22}-\lambda B_{22} & A_{23}-\lambda B_{23} \\ 0 & 0 & 0 & A_{32}-\lambda B_{32} & A_{33}-\lambda B_{33} \\ 0 & 0 & 0 & D_2 & D_3 \end{bmatrix}.$$

We claim the pencil on the right above is in GUPTRI form, so that its reducing subspaces are easy to discern. This is because the $[C_{11}|A_{11} + C_1K - \lambda B_{11}]$ block, because of the nonsingularity of B, can only have L_k blocks and finite eigenvalues in its KCF, and the other diagonal block can only have L_j^T blocks and infinite eigenvalues in its KCF. Therefore the largest left and right reducing subspaces of (13) not including any infinite eigenvalues are given by

$$\mathbf{P}_i = \mathbf{R}\left(\begin{bmatrix} P_1 \\ 0 \end{bmatrix}\right) \quad \text{and} \quad \mathbf{Q}_i = \mathbf{R}\left(\begin{bmatrix} S_1 & K \\ 0 & Q_1 \end{bmatrix}\right)$$

proving claim (i) of the theorem.

Claims (ii) and (iii) follow by applying Theorem 1 to the submatrix

$$[C_{11}|A_{11} + C_1K - \lambda B_{11}]. \qquad \square$$

In order to use our perturbation theory for reducing subspaces to get error bounds for $\mathbf{V}^*$ and $\mathbf{R}^*$, we must use the left reducing subspace instead of the right, because the right subspace is not the direct sum of a constant space and $\mathbf{V}^*$ (or $\mathbf{R}^*$). If $B = I$, then we can use the left subspace directly, otherwise we must modify the bounds by the following lemma. We state the lemma in a general way so as to cover both $\mathbf{V}^*$ and $\mathbf{R}^*$ simultaneously. Thus for $\mathbf{V}^*$ choose $\mathbf{S} = \mathbf{R}(P_{i1})$ (see Theorem 3(i)), $H_S = B^{-1}$, $\mathbf{T}$ as the perturbed $\mathbf{R}(P_{i1})$, η as the bound on $\theta_{\max}(\mathbf{S}, \mathbf{T})$ to be provided by the algorithm, H_T as the approximation to B^{-1} used to compute $\mathbf{V}^* = B^{-1}\mathbf{T}$, and δ as the bound on the error in B^{-1} arising from round-off and the maximum perturbation made by the algorithm; $\mathbf{R}^*$ is similar.

LEMMA 2. *If* $\tan\theta_{\max}(\mathbf{S}, \mathbf{T}) \leqq \eta < \pi/2$ *and* $\|H_S - H_T\| < \delta$, *then*

$$\tan\theta_{\max}(H_S\mathbf{S}, H_T\mathbf{T}) \leqq \left[\kappa(H_S)\eta + \frac{\delta(1+\eta)}{\sigma_{\min}(H_S)}\right] \Big/ \left[1 - \kappa(H_S)\eta - \frac{\delta(1+\eta)}{\sigma_{\min}(H_S)}\right]$$

if the denominator is positive.

Proof. Let $P = [P_1|P_2]$ be a unitary matrix whose first dim (**S**) columns span **S**. Letting the columns of P be a new basis for our space, we see that without loss of generality we can assume that

$$\mathbf{S} = \mathbf{R}\left(\begin{bmatrix} I \\ 0 \end{bmatrix}\right) \quad \text{and} \quad \mathbf{T} = \mathbf{R}\left(\begin{bmatrix} I \\ Z \end{bmatrix}\right)$$

where $\|Z\| \leqq \eta$. Let $Q = [Q_1|Q_2]$ be a unitary matrix whose first dim (**S**) columns Q_1 span $\mathbf{R}(H_SP_1)$. Then denoting

$$Q^*H_SP = \begin{bmatrix} H_{S11} & H_{S12} \\ 0 & H_{S22} \end{bmatrix} \quad \text{and} \quad Q^*H_TP = \begin{bmatrix} H_{T11} & H_{T12} \\ H_{T21} & H_{T22} \end{bmatrix}$$

we see that

$$H_S\mathbf{S} = Q\mathbf{R}\left(\begin{bmatrix} H_{S11} \\ 0 \end{bmatrix}\right) \quad \text{and} \quad H_T\mathbf{T} = Q\mathbf{R}\left(\begin{bmatrix} H_{T11} + H_{T12}Z \\ H_{T21} + H_{T22}Z \end{bmatrix}\right)$$

so that

$$\theta_{\max}(H_S\mathbf{S}, H_T\mathbf{T}) = \theta_{\max}\left(\mathbf{R}\left(\begin{bmatrix} H_{S11} \\ 0 \end{bmatrix}\right), \mathbf{R}\left(\begin{bmatrix} H_{T11} + H_{T12}Z \\ H_{T21} + H_{T22}Z \end{bmatrix}\right)\right)$$

$$= \theta_{\max}\left(\mathbf{R}\left(\begin{bmatrix} I \\ 0 \end{bmatrix}\right), \mathbf{R}\left(\begin{bmatrix} I \\ (H_{T21} + H_{T22}Z)\cdot(H_{T11} + H_{T12}Z)^{-1} \end{bmatrix}\right)\right)$$

$$= \arctan \|(H_{T21} + H_{T22}Z)\cdot(H_{T11} + H_{T12}Z)^{-1}\|$$

$$\leqq \arctan\left(\kappa(H_S)\eta + \frac{\delta(1+\eta)}{\sigma_{\min}(H_S)} \Big/ 1 - \kappa(H_S)\eta - \frac{\delta(1+\eta)}{\sigma_{\min}(H_S)}\right). \qquad \square$$

Note that the last bound in the proof may seriously overestimate the next to last bound. The next to last bound should be used if sharper bounds are desired.

Finally, we turn to invariant zeros of the complete system

$$\begin{aligned} B\dot{x} &= Ax + Cu, \\ y &= Dx + Fu. \end{aligned} \tag{14}$$

As defined in [16], the invariant zeros are the finite Smith zeros of the pencil

$$\begin{bmatrix} C & A - \lambda B \\ F & D \end{bmatrix} \tag{15}$$

which are nothing more than the finite eigenvalues of (15) [7]. Therefore, the perturbation theory for generalized eigenvalues of pencils in the next section provides bounds for invariant zeros.

5. Perturbation theory for reducing subspaces and generalized eigenvalues. In this section we present computable error bounds for reducing subspaces and eigenvalues of matrix pencils. We assume we have reduced the m-by-n pencil to the GUPTRI form (5)

$$P^{-1}(H - \lambda G)Q = \begin{bmatrix} H_{11} - \lambda G_{11} & H_{12} - \lambda G_{12} \\ 0 & H_{22} - \lambda G_{22} \end{bmatrix}$$

where P and Q are unitary, $H_{ii} - \lambda G_{ii}$ is m_i by n_i, $H_{11} - \lambda G_{11}$ has only L_j blocks and a regular part with spectrum σ_1 in its KCF, $H_{22} - \lambda G_{22}$ has only L_j^T blocks and a regular part with spectrum σ_2 in its KCF, and σ_1 and σ_2 are disjoint. Recall that in this coordinate system, the left and right reducing subspaces are spanned by $[I_{m_1}|0]^T$ and $[I_{n_1}|0]^T$. Algorithms for reducing general pencils to this form are described in [8], [9], [15], [17], [20]. Our algorithm is approximately twice as fast as Van Dooren's [15] on general pencils and will be described in another paper. In the case of the special pencils (7), (10), (13), and (15) in the last section, more efficient algorithms for the case $G = I$ appear in [6], [16].

To present our bounds we need some definitions. Details and proofs may be found in [4]. We first need to blockdiagonalize (5), which means solving the equation

$$\begin{bmatrix} I_{m_1} & -L \\ 0 & I_{m_2} \end{bmatrix} \cdot \begin{bmatrix} H_{11} - \lambda G_{11} & H_{12} - \lambda G_{12} \\ 0 & H_{22} - \lambda G_{22} \end{bmatrix} \cdot \begin{bmatrix} I_{n_1} & R \\ 0 & I_{n_2} \end{bmatrix} = \begin{bmatrix} H_{11} - \lambda G_{11} & 0 \\ 0 & H_{22} - \lambda G_{22} \end{bmatrix}$$

for L and R, or

$$\begin{aligned} H_{11}R - LH_{22} &= -H_{12}, \\ G_{11}R - LG_{22} &= -G_{12} \end{aligned}$$

which is a generalized form of Sylvester's equation. We can rewrite this in terms of Kronecker products as follows:

$$\begin{bmatrix} I_{n_2} \otimes H_{11} & -H_{22}^T \otimes I_{m_1} \\ I_{n_2} \otimes G_{11} & -G_{22}^T \otimes I_{m_1} \end{bmatrix} \cdot \begin{bmatrix} \text{col } R \\ \text{col } L \end{bmatrix} \equiv Z_u \cdot \begin{bmatrix} \text{col } R \\ \text{col } L \end{bmatrix} = \begin{bmatrix} -\text{col } H_{12} \\ -\text{col } G_{12} \end{bmatrix}.$$

This is a set of $2m_1n_2$ linear equations in $n_1n_2 + m_1m_2$ unknowns, the entries of L and R. Since $m_1 \leqq n_1$ and $m_2 \geqq n_2$, we see we have at least as many unknowns as equations with equality if and only if $H - \lambda G$ is regular. When $H - \lambda G$ is singular, Z_u has full row rank and so there is a (nonunique) solution. We choose the minimum norm solutions L_0 and R_0 because this gives us the best bound later. Let $p \equiv (1 + \|L_0\|^2)^{1/2}$ and $q \equiv (1 + \|R_0\|^2)^{1/2}$. p and q play the same role for this theory as the norm of the projection does for the standard eigenproblem: they measure the sensitivity of eigenspaces (and eigenvalues) with respect to changes in H and G. Indeed, if $G = I$, they are equal to the norm of a projection onto an invariant subspace of H. For the generalized eigenproblem we need both a left and a right projection norm since the left and right spaces may differ.

We will also need the quantity

$$\mathrm{Dif}_u(H_{11}, H_{22}; G_{11}, G_{22}) \equiv \sigma_{\min}(Z_u)$$

which is nonzero since Z_u has full rank. Similarly, we need

$$\mathrm{Dif}_l(H_{11}, H_{22}; G_{11}, G_{22}) \equiv \sigma_{\min}\left(\begin{bmatrix} H_{11}^T \otimes I_{m_2} - I_{n_1} \otimes H_{22} \\ G_{11}^T \otimes I_{m_2} - I_{n_1} \otimes G_{22} \end{bmatrix}\right)$$

which is nonzero if and only if Dif_u is nonzero. We can show that all these definitions are really coordinate free allowing us to write $\mathrm{Dif}_l(\sigma_1, \sigma_2)$ ($\mathrm{Dif}_u(\sigma_1, \sigma_2)$) when $H - \lambda G$ is known from context or just Dif_l (Dif_u) if σ_1 is known as well.

Both Dif_l and Dif_u measure how close the KCFs of the $H_{11} - \lambda G_{11}$ and $H_{22} - \lambda G_{22}$ are to one another. They generalize the operator

$$\mathrm{sep}(H_{11}, H_{22}) \equiv \sigma_{\min}(I_{n_2} \otimes H_{11} - H_{22}^T \otimes I_{n_1})$$

which measures the separation of the spectra of two square matrices H_{11} and H_{22} [11]: it (under)estimates the size of the smallest perturbation needed to make H_{11} and H_{22} have a common eigenvalue. Indeed, when $G = I$, Dif_l, Dif_u, and sep are all almost equal.

Now we may state the following.

THEOREM 4. *Let $H - \lambda G$ be an m-by-n singular pencil of the form* (5). *Let* **P** *and* **Q** *be the left and right reducing subspaces of $H - \lambda G$ belonging to σ_1. Let them have dimensions m_1 and n_1, respectively. Let $\hat{m} \equiv \min(m_1, m - m_1)$ and $\hat{n} \equiv \min(n_1, n - n_1)$. Define*

$$\Delta \equiv \frac{\min(\mathrm{Dif}_u(\sigma_1, \sigma_2), \mathrm{Dif}_l(\sigma_1, \sigma_2))}{(p^2 + q^2)^{1/2} + 2 \cdot \max(p, q)}.$$

Then if $(H + \delta H) - \lambda(G + \delta G)$ has reducing subspaces $\mathbf{P}_\delta$ and $\mathbf{Q}_\delta$ of the same dimensions as **P** *and* **Q**, *respectively, and*

$$x \equiv \frac{\|(\delta H, \delta G)\|_E}{\Delta} < 1, \tag{16}$$

then one of the following two cases must hold:

Case 1.

$$\theta_{\max}(\mathbf{P}, \mathbf{P}_\delta) \leqq \arctan\left(\frac{x}{p - x \cdot (p^2 - 1)^{1/2}}\right) \leqq \arctan(x \cdot (p + (p^2 - 1)^{1/2}))$$

and

$$\theta_{\max}(\mathbf{Q}, \mathbf{Q}_\delta) \leqq \arctan\left(\frac{x}{q - x \cdot (q^2 - 1)^{1/2}}\right) \leqq \arctan(x \cdot (q + (q^2 - 1)^{1/2})).$$

In other words, both angles are small, bounded above by a multiple of the norm of the perturbation $\|(\delta H, \delta G)\|_E$.

Case 2. *Either*

$$\theta_{\max}(\mathbf{P}, \mathbf{P}_\delta) \geqq \arctan\left(\frac{1}{\sqrt{2\hat{m}\cdot p + (p^2-1)^{1/2}}}\right)$$

or

$$\theta_{\max}(\mathbf{Q}, \mathbf{Q}_\delta) \geqq \arctan\left(\frac{1}{\sqrt{2\hat{n}\cdot q + (q^2-1)^{1/2}}}\right).$$

In other words, at least one of the angles between perturbed and unperturbed reducing subspaces is bounded away from 0.

The significance of the criterion (16) is as follows. Since the denominator of Δ is the "speed" with which the KCFs of the two diagonal blocks of $H - \lambda G$ can change, and the numerator is the "distance" between their KCFs, (16) states that $\|(\delta H, \delta G)\|_E$ is small enough so that the assumptions about the pencil being reducible to GUPTRI form (5) hold under perturbations. In the special case when $G = I$, Δ can be shown to reduce to a known good estimate of the largest perturbation before an eigenvalue from σ_1 coalesces with an eigenvalue from σ_2 causing their respective invariant subspaces to overlap and any perturbation theory to break down [3], [11]. In other words, we can only do perturbation theory for a certain feature until the perturbation becomes so large we can no longer guarantee that the feature is well defined.

The theorem has two cases because of a "labeling" problem. A singular pencil may have several reducing subspaces of the same dimension, just as a matrix may have several invariant subspaces of the same dimension, each corresponding to a different set of eigenvalues. In the case of the matrix, we can "label" each invariant subspace with the eigenvalues to which it belongs, and identify a perturbed subspace by its perturbed eigenvalues. Thus a perturbation theorem for invariant subspaces would read "a small perturbation in the matrix perturbs the eigenvalues in σ_1 to a nearby set σ_1', and the invariant subspace of the perturbed matrix corresponding to σ_1' is close to the unperturbed invariant subspace corresponding to σ_1." The analogous theorem for singular pencils must be stated differently, because the perturbed pencil may have no eigenvalues at all to use as labels. Therefore we must say that if it has a reducing subspace of the right dimension, this must either be a small perturbation of the original unperturbed one (Case 1) or a different one (Case 2). In fact, applying Theorem 4 to square pencils of the form $H - \lambda I$, we can interpret it as providing perturbation bounds for the invariant subspace belonging to σ_1' in Case 1 and for all other invariant subspaces of the same dimension belonging to any $\sigma_2' \neq \sigma_1'$ in Case 2.

In practice, deciding which case applies is no problem, since the reduction algorithm will try to pick the same one each time, so that Case 1 applies. For example, when computing the controllable subspace the computed reducing subspace is always the minimal one.

A proof of Theorem 4 may be found in [4, Thm. 5].

In practice, the theorem may be applied as follows. Consider Fig. 1. The original input pencil is $H - \lambda G$; it lies on or near a surface S of pencils of some fixed Kronecker structure (e.g., those pencils representing control systems with two uncontrollable modes). The user supplies to the algorithm both $H - \lambda G$ and an upper bound δ on its distance to the surface. δ may be the user's best estimate of the noise in his data, or a stability margin in case he wants to know if his system is close to one with a particular structure. δ should be at least a modest multiple of the machine precision. If $H - \lambda G$ is close

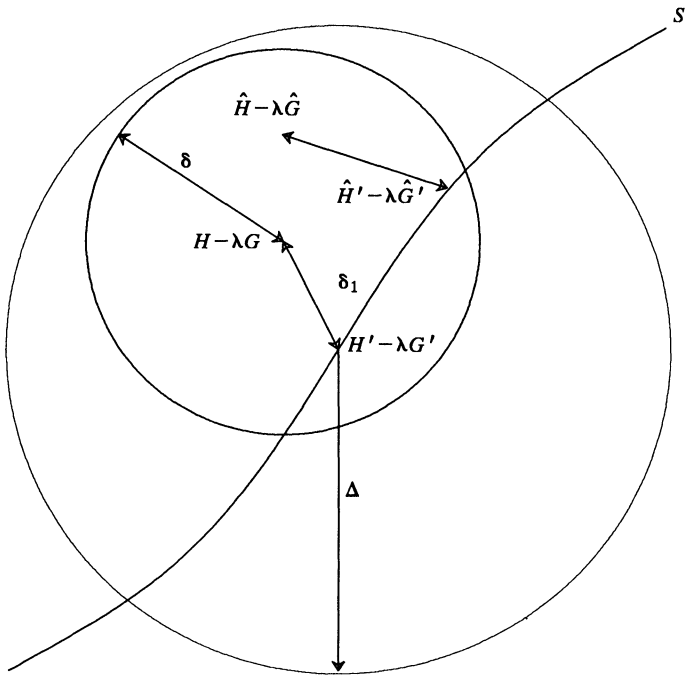


FIG. 1. *Perturbation theory for singular pencils.*

enough to S, the algorithm will find a nearby system $H' - \lambda G'$ lying directly on S, and compute its decomposition (5) (see Fig. 1). $\delta_1 \leqq \delta$ is the actual distance

$$\|(H-H', G-G')\|_E$$

between the two pencils.

Now we may compute **P**, **Q**, Δ, p, and q of the theorem. The perturbation bounds in Theorem 4 apply to all pencils on the surface and within distance Δ of $H' - \lambda G'$. We may test the theorem as follows. We take $H - \lambda G$ and add random noise of size at most δ to each component to get a perturbed pencil $\hat{H} - \lambda\hat{G}$. We input this pencil to the algorithm. In general the algorithm will compute an $\hat{H}' - \lambda\hat{G}'$ on S (see Fig. 1). If $\hat{H}' - \lambda\hat{G}'$ is within distance Δ of $H' - \lambda G'$ we compute its reducing subspaces, measure their actual angles from **P** and **Q**, and see if these are either less than their upper bounds or greater than their upper bounds in the theorem. We report on experiments of this type in the next section.

Now we turn to eigenvalue bounds. We deal first with the simple case in which there is only one type of singular structure in the KCF: L_j blocks or L_j^T blocks.

THEOREM 5. *Suppose that Case* 1 *of Theorem* 4 *holds. Suppose further that the block* $H_{22} - \lambda G_{22}$ *is regular. (This implies* $H - \lambda G$ *has no* L_j^T *blocks in its* KCF.) *Then the spectrum of the perturbed pencil* $(H + \delta H) - \lambda(G + \delta G)$ *includes the spectrum of*

$$(H_{22} + \delta H'_{22}) - \lambda(G_{22} + \delta G'_{22})$$

where

$$\|(\delta H'_{22}, \delta G'_{22})\|_E \leqq \sqrt{2} \cdot q \cdot \|(\delta H, \delta G)\|_E.$$

Similarly, if we instead assume $H_{11} - \gamma G_{11}$ *is regular, then the spectrum of the perturbed pencil* $(H + \delta H) - \lambda(G + \delta G)$ *includes the spectrum of*

$$(H_{11} + \delta H'_{11}) - \lambda(G_{11} + \delta G'_{11})$$

where

$$\|(\delta H'_{11}, \delta G'_{11})\|_E \leqq \sqrt{2} \cdot p \cdot \|(\delta H, \delta G)\|_E.$$

A proof may be found in [4, Thm. 6].

Thus we have reduced the problem to one of perturbation theory for eigenvalues of regular pencils, a well-studied area [1], [4], [11], [12], [13], [14], [20]. For simplicity, the bounds we implemented in our code, a generalization of and improvement on the Bauer–Fike Theorem, assume all the eigenvalues are simple, although this could easily be changed.

It remains to show how to do perturbation theory for eigenvalues of pencils with both kinds of singular blocks in their KCFs. We must simply reduce to GUPTRI form twice, once to isolate the L_j^T blocks, and the second time to apply Theorem 5.

COROLLARY 1. *Suppose* $H - \lambda G$ *has* L_j *blocks,* L_j^T *blocks, and a regular part in its* KCF. *Let*

$$\begin{bmatrix} H_{11} - \lambda G_{11} & H_{12} - \lambda G_{12} \\ 0 & H_{22} - \lambda G_{22} \end{bmatrix}$$

be the GUPTRI *form of* $H - \lambda G$ *where* $H_{22} - \lambda G_{22}$ *contains all the* L_j^T *blocks and* $H_{11} - \lambda G_{11}$ *contains all the* L_j *blocks and the regular part. Let* Δ_1, p_1, *and* q_1 *be the quantities of Theorem* 4 *associated with this decomposition. Let*

$$\begin{bmatrix} H'_{11} - \lambda G'_{11} & H'_{12} - \lambda G'_{12} \\ 0 & H'_{22} - \lambda G'_{22} \end{bmatrix}$$

be the GUPTRI *form of* $H_{11} - \lambda G_{11}$ *where* $H'_{22} - \lambda G'_{22}$ *is regular and* $H'_{11} - \lambda G'_{11}$ *has only* L_j *blocks in its* KCF. *Let* Δ_2, p_2, and q_2 *be the quantities of Theorem* 4 *associated with this decomposition.*

Then if the perturbed pencil $(H + \delta H) - \lambda(G + \delta G)$ *has the same size right singular, regular, and left singular blocks as* $H - \lambda G$, *and*

$$\|(\delta H, \delta G)\|_E \leqq \min\left(\Delta_1, \frac{\Delta_2}{\sqrt{2}p_1}\right),$$

then $(H + \delta H) - \lambda(G + \delta G)$ *has eigenvalues equal to the eigenvalues of*

$$(H'_{22} + \delta H') - \lambda(G'_{22} + \delta G')$$

with

$$\|(\delta H', \delta G')\|_E \leqq 2 \cdot q_2 \cdot p_1 \cdot \|(\delta H, \delta G)\|_E.$$

It should be clear how to modify this corollary if we want bounds assuming only some of the eigenvalues are preserved by perturbations.

6. Numerical examples. In this section we will report on numerical experiments using our algorithm for reduction to GUPTRI form and the perturbation bounds. All our tests were made using the following scheme.

(1) Choose a nongeneric pencil $H - \lambda G$ and a "rule" for choosing a particular set of reducing subspaces. (For example, for controllable subspaces we choose $H - \lambda G =$

$[C|A - \lambda B]$ and minimal reducing subspaces as in Theorem 1.) Using the GUPTRI algorithm compute its reducing subspaces $\mathbf{P}$ and $\mathbf{Q}$ and the quantities Δ, p, and q of Theorem 4. Also compute the eigenvalues and eigenvalue bounds of Corollary 1 if desired.

(2) Add random noise of size ε_n to $H - \lambda G$ to get a perturbed pencil $\hat{H} - \lambda \hat{G}$. Input $\hat{H} - \lambda \hat{G}$ to the GUPTRI algorithm along with a bound ε_u on the distance the algorithm may perturb $\hat{H} - \lambda \hat{G}$. Let $\hat{H}' - \lambda \hat{G}'$ denote the output pencil in GUPTRI form.

(3) Compute the reducing subspaces $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ of $\hat{H}' - \lambda \hat{G}'$ according to the rule chosen in step (1). If $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ have the same dimensions as $\mathbf{P}$ and $\mathbf{Q}$, and if $\|(H - \hat{H}', G - \hat{G}')\|_E < \Delta$ as required in the hypotheses of Theorem 4, test to see if either Case 1 or Case 2 holds. If there are eigenvalues, test to see if they are as close as predicted by Corollary 1.

(4) Repeat steps (2) and (3) for different noise sizes ε_n, different random noise, and different bounds ε_u.

We tested nine cases using this scheme, three for controllable subspaces and uncontrollable modes (using Theorems 1 and 4 and Corollary 1), three for $\mathbf{V}^*$ (using Theorems 3 and 4), and three for $\mathbf{R}^*$ (using Theorems 3 and 4). In all cases $B = I$. For each case we tried all 24 combinations of ε_n chosen from $10^{-10}, 10^{-9}, \cdots, 10^{-3}$ and ε_u chosen from $10^{-7}, 10^{-5}, 10^{-3}$ (the details of how ε_u is used imply that the effective ε_u may be up to a factor of 1000 smaller). For each choice of case, ε_n and ε_u, 10 random pencils were tried, for a total of 2160 pencils.

We collected statistics on how often the hypotheses of Theorem 4 were satisfied or why they were not satisfied, how often Cases 1 or 2 arose if they were satisfied, and how good our upper bounds were in Case 1. Right subspaces were used in all cases.

In summary, the results agreed with the predictions of the perturbation theory. In almost all cases either Case 2 of Theorem 4 held or a reducing subspace of a different dimension (usually a generic one) was computed. Case 2 of Theorem 4 held when ε_u (the estimate of the size of the noise supplied to the algorithm) sufficiently exceeded ε_n (the actual size of the noise). How much ε_u had to exceed ε_n depended on the conditioning of the problem.

The three cases chosen to compute controllable subspaces were

$$\left|\begin{matrix}1\\1\\0\\0\end{matrix}\right|\left|\begin{matrix}-2 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 2\end{matrix}\right|, \quad \left|\begin{matrix}1\\1\\0\\0\end{matrix}\right|\left|\begin{matrix}-2 & -10 & 0 & 0\\ 0 & -1 & 0 & 0\\ 0 & 0 & 1 & -7.5\\ 0 & 0 & 0 & 2\end{matrix}\right|, \quad \left|\begin{matrix}1\\1\\0\\0\end{matrix}\right|\left|\begin{matrix}-2 & -100 & 0 & 0\\ 0 & -1 & 0 & 0\\ 0 & 0 & 1 & -75\\ 0 & 0 & 0 & 2\end{matrix}\right|$$

where we use the notation of (7). We call these examples C1, C2, and C3. We also made a random orthogonal change of coordinates on each one. Note that each A matrix has successively more ill-conditioned eigenvalues, as seen by the size of the off-diagonal elements. Each one has a two-dimensional controllable subspace and uncontrollable modes at 1 and 2.

The results are summarized in Table 1. We expected that as long as ε_u (the estimate supplied to the algorithm of the maximum noise in the data) exceeded ε_n (the actual noise) sufficiently, the algorithm would compute a controllable subspace of dimension 2, and otherwise a larger one. This was generally true, with the larger one being generic in almost all cases. However, C3 needed to have $\varepsilon_u/\varepsilon_n$ much larger than C2, and C2 needed $\varepsilon_u/\varepsilon_n$ much larger than C1 to compute a two-dimensional space, as can be seen by the decreasing proportion of trials corresponding to Case 1 of Theorem 4. This is apparently a result of the increasing sensitivity of the eigenproblem of A. Also, the quality of our upper bounds decreased as this sensitivity increased, as evidenced by the increasing ratios $\theta_{\text{bnd}}/\theta_{\text{true}}$ of our upper bound on the angle (between unperturbed and perturbed

TABLE 1
Results of computing controllable subspaces.

	C1	C2	C3
Case 1 of Theorem 4	58%	39%	27%
$\|(H - \hat{H}', G - \hat{G}')\|_E > \Delta$	0%	0%	1%
dimension different	42%	61%	72%
avg $(\theta_{bnd}/\theta_{true})$	5.	27.	132.
max $(\theta_{bnd}/\theta_{true})$	17.	36.	1250.

spaces) to the true angle. In fact, C3 was sufficiently ill-conditioned that when ε_u was 10^{-3}, the algorithm found a different nearby uncontrollable system with almost the same controllable space but quite different uncontrollable modes: 0 and 3. Another measure of this increasing ill-condition is the ratio $\|(H - \hat{H}', G - \hat{G}')\|_E / \|(H - \hat{H}, G - \hat{G})\|_E$ of the distance between the original system and the output of the algorithm to the distance between the original system and the input to the algorithm; if the system is well behaved this ratio should not exceed 1 by much, indicating that the algorithm can project the perturbed system nearly perpendicularly back onto the surface of systems with the original system's structure. The maximum value of this ratio was 16 for C1, 121 for C2, and 885 for C3. Nonetheless, our bounds were generally realistic, generally not exceeding the true perturbations by a very large factor.

The results of computing the uncontrollable modes are shown in Table 2. Here e_{bnd} is the bound on the perturbation in the eigenvalue, and e_{dif} is the true perturbation, where we measure perturbations as "angles": an eigenvalue e_i is written as $\tan \theta_i$ and we measure the difference between e_1 and e_2 by $|\theta_1 - \theta_2|$. This is related to the chordal metric [12]. These results seem much poorer than the ones in Table 1 until we examine the bounds themselves: they are quite small in absolute value, and closer inspection shows that if all the random noise were added to the regular part of the pencil, the bound e_{bnd} could be nearly achieved. For some reason, the random noise seems to affect the controllable subspaces much more than the uncontrollable modes.

The three cases chosen for computing $\mathbf{V}^*$ and $\mathbf{R}^*$ were

$$\left[\begin{array}{ccc|cccc} 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right], \quad \left[\begin{array}{cc|ccccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right],$$

$$\left[\begin{array}{cc|ccccc} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{array}\right].$$

TABLE 2
Results of computing uncontrollable modes.

	C1	C2	C3
avg (e_{bnd}/e_{dif})	311.	1171.	2960.
max (e_{bnd}/e_{dif})	6180.	20280.	176000.

TABLE 3
Properties of unperturbed $\mathbf{V}^*$ *and* $\mathbf{R}^*$.

	AB1	AB2	AB3
dim (ker D)	3	3	2
dim ($\mathbf{V}^*$)	3 (generic)	1 (nongeneric)	1 (nongeneric)
dim ($\mathbf{R}^*$)	2 (nongeneric)	0 (generic)	0 (generic)

TABLE 4
Results of computing $\mathbf{V}^*$ *and* $\mathbf{R}^*$.

	AB1 $\mathbf{V}^*$	AB1 $\mathbf{R}^*$	AB2 $\mathbf{V}^*$	AB3 $\mathbf{V}^*$
Case 1 of Theorem 4	100%	62%	73%	73%
dimension different	0%	38%	27%	27%
avg ($\theta_{bnd}/\theta_{true}$)	54.	39.	26.	56.
max ($\theta_{bnd}/\theta_{true}$)	210.	73.	53.	106.

where we use the notation of (13). We call these examples AB1, AB2, and AB3. We also made a random orthogonal change of coordinates. The properties of $\mathbf{V}^*$ and $\mathbf{R}^*$ for these examples are summarized in Table 3. The results of the test runs are given in Table 4. Results for $\mathbf{R}^*$ for AB2 and AB3 are not shown; the generic $\mathbf{R}^* = \{0\}$ was computed for all ε_n and ε_u. In computing $\mathbf{V}^*$ for perturbed AB1, the generic $\mathbf{V}^* = \ker D$ was also computed for all ε_n and ε_u, and Case 1 of Theorem 4 always applied. In the other three cases shown, roughly speaking as long as ε_u exceeded ε_n Case 1 of Theorem 4 occurred; otherwise the perturbed $\mathbf{V}^*$ or $\mathbf{R}^*$ had a different dimension than the unperturbed one (for $\mathbf{R}^*$ and AB1, we needed $\varepsilon_u \geqq 10\varepsilon_n$). These $\mathbf{V}^*$ and $\mathbf{R}^*$ of different dimensions corresponded to generic systems in all but one percent of the experiments for AB3, when an originally infinite eigenvalue became a very large finite one (this situation would have been avoided using an algorithm specialized for computing $\mathbf{V}^*$ [16]). The ratio $\theta_{bnd}/\theta_{true}$ of the bound on the perturbation in $\mathbf{V}^*$ or $\mathbf{R}^*$ to the true perturbation in Theorem 4 was almost always less than 100, and had a maximum value of 210. Also the ratio

$$\|(H-\hat{H}', G-\hat{G}')\|_E / \|(H-\hat{H}, G-\hat{G})\|_E$$

never exceeded four for any example.

We believe that these results would improve if specialized algorithms [16] which respect the structures in (6) and (10) were used instead of a general purpose algorithm.

REFERENCES

[1] K. E. CHU, *Exclusion theorems and the perturbation analysis of the generalized eigenvalue problem*, Numerical Analysis Report NA/11/85, Mathematics Dept., University of Reading, Reading, England.

[2] D. COBB, *Controllability, observability, and duality in singular systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1076–1082.

[3] J. DEMMEL, *Computing stable eigendecompositions of matrices*, Linear Algebra Appl., 79 (1986), pp. 163–193.

[4] J. DEMMEL AND B. KÅGSTRÖM, *Stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88/89 (1987), pp. 137–186.

[5] ———, *Stably computing the Kronecker structure and reducing subspaces of singular pencils* $A - \lambda B$ *for uncertain data*, in Proc. of Conference on Large Eigenvalue Problems, J. Cullum and R. Willoughby, eds., IBM Europe Institute, Oberlech, Austria, July 1985, North Holland, Amsterdam, 1986.

[6] A. EMAMI-NAEINI AND P. VAN DOOREN, *Computation of zeros of linear multivariable systems*, Automatica, 18 (1982), pp. 415–430.

[7] F. GANTMACHER, *The Theory of Matrices*, Vol. II (Transl.), Chelsea, New York, 1959.

[8] B. KÅGSTRÖM, *The generalized singular value decomposition and the general* $(A - \lambda B)$ *problem*, BIT, 24 (1984), pp. 568–583.

[9] ———, *RGSVD—An algorithm for computing the Kronecker structure and reducing subspaces of singular matrix pencils* $A - \lambda B$, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 185–211.

[10] B. KÅGSTRÖM AND A. RUHE, eds., *Matrix Pencils*, Proc. Pite Havsbad, 1982, Lecture Notes in Mathematics, 973, Springer-Verlag, New York, Berlin, 1983.

[11] G. STEWART, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 752–764.

[12] ———, *Gershgorin theory for the generalized eigenproblem* $Ax = \lambda Bx$, Math. Comp., 29 (1975), pp. 600–606.

[13] ———, *On the sensitivity of the eigenvalue problem* $Ax = \lambda Bx$, SIAM J. Numer. Anal., 9 (1972), pp. 669–686.

[14] J-G. SUN, *Perturbation analysis for the generalized eigenvalue and generalized singular value problem*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 221–244.

[15] P. VAN DOOREN, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–141.

[16] ———, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 111–128.

[17] ———, *Reducing subspaces: definitions, properties and algorithms*, in Matrix Pencils, Proc. Pite Havsbad, 1982, Lecture Notes in Mathematics, 973, Springer-Verlag, New York, Berlin, 1983, pp. 58–73.

[18] ———, *Reducing subspaces: computational aspects and applications in linear systems theory*, Proc. 5th Internat. Conference on Analysis and Optimization of Systems, Versailles, Lecture Notes on Control and Information Science, 44, Springer-Verlag, New York, Berlin, 1982.

[19] G. VERGHESE, *Infinite frequency behaviour in generalized dynamical systems*, Ph.D. dissertation, Information Systems Laboratory, Stanford University, Stanford, CA, 1978.

[20] J. WILKINSON, *Linear differential equations and Kronecker's canonical form*, in Recent Advances in Numerical Analysis, C. de Boor and G. Golub, eds., Academic Press, New York, 1978, pp. 231–265.

[21] M. WONHAM, *Linear Multivariable Theory. A Geometric Approach*, 2nd ed., Springer-Verlag, New York, Berlin, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

A NOTE ON THE SHORTED OPERATOR*

C. A. BUTLER† AND T. D. MORLEY‡

Abstract. The Schur complement of a partitioned operator

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

is defined by the formula $S(A) = A_{11} - A_{12}A_{22}^{-1}A_{21}$. In finite dimensions $S(A)$ is the unique map $c \mapsto d$ defined by the equations $A_{11}c + A_{12}y = d$, $A_{21}c + A_{22}y = 0$. In infinite dimensions the shorted operator of a positive operator generalizes the Schur complement; however, the above matrix equations no longer hold. We show in what sense the above equations approximately hold. Applications to infinite networks are shown.

Key words. Schur complement, electrical networks, shorted operator

AMS(MOS) subject classifications. 15A09, 15A30, 47A99, 94A20

1. Introduction: Finite dimensions. Let A be a bounded linear operator on a Hilbert space $\mathscr{H}$. Let S be a closed subspace of $\mathscr{H}$. Then with respect to a suitable orthonormal basis we may write

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

with A_{11}: $S \to S$, A_{12}: $S^{\perp} \to S$, A_{21}: $S \to S^{\perp}$, and A_{22}: $S^{\perp} \to S^{\perp}$.

The Schur complement of A to a subspace S is the operator $S(A)$: $S \to S$ defined by the formula

$$S(A) = A_{11} - A_{12}A_{22}^{-1}A_{21}. \tag{1}$$

Of course, the Schur complement is not defined unless $(A_{22})^{-1}$ exists.

It is easy to see that the Schur complement, if it exists, is uniquely defined by the matrix equation

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} c \\ w \end{bmatrix} = \begin{bmatrix} S(A)c \\ 0 \end{bmatrix}$$

where, of course, w depends linearly on c.

If $\mathscr{H}$ is finite-dimensional and if the matrix A is positive, i.e., if $A = A^*$ and $(Ax, x) \geqq 0$ for all x, then the above matrix equation defines a unique operator $S(A)$ irrespective of the invertibility of A_{22}. In this case, we refer to $S(A)$ as the shorted operator. The shorted operator was introduced by Anderson [1] in connection with electrical networks. We now briefly describe this connection.

An n-port is an electrical device with n pairs of terminals to the outside world. Its external behavior is determined by a matrix $\{a_{ij}\}$. If a current source of x_j amps is connected across the jth terminal pair (or *port*), the voltage v_i across the ith terminal is given by

$$v_i = \sum_{j=1}^{n} a_{ij}x_j,$$

or in matrix notation

$$\mathbf{v} = A\mathbf{x}.$$

The matrix A is termed the impedance matrix of the n-port.

* Received by the editors January 17, 1986; accepted for publication (in revised form) April 10, 1987.
† Department of Mathematics, Georgia College, Milledgeville, Georgia 31061.
‡ School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

If the n-port consists entirely of resistors and (ideal) transformers, then the matrix A will be positive; see Fig. 1. (Moreover any positive matrix arises in this way [20].)

If the last several ports of an n-port are shorted together, as in Fig. 2, then any current is possible across the shorted terminals. However, the voltage across the shorted terminals must be zero. This gives rise to the matrix equations

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} c \\ w \end{bmatrix} = \begin{bmatrix} v \\ 0 \end{bmatrix}. \tag{2}$$

Solving formally for v as a function of c, we have

$$(A_{11} - A_{12}A_{22}^{-1}A_{21})c = v.$$

Thus the Schur complement (or shorted operator) represents the impedance matrix of the shorted n-port.

Anderson has shown [1] that if A is a positive matrix, then (2) uniquely defines the impedance matrix of the shorted n-port.

The following definition is equivalent to that given by Ando [4].

DEFINITION. An operator A is termed complementable to S if given any $c \in S$, there is a $w \in S^{\perp}$ and a unique $d \in S$ such that

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} c \\ w \end{bmatrix} = \begin{bmatrix} d \\ 0 \end{bmatrix}.$$

If A is complementable, we define the generalized Schur complement by $S(A)c = d$ in the above equations.

Thus in finite dimensions positive operators are complementable, and the shorted operator construction of Anderson [1] and the generalized Schur complement of Ando [4] agree for positive operators.

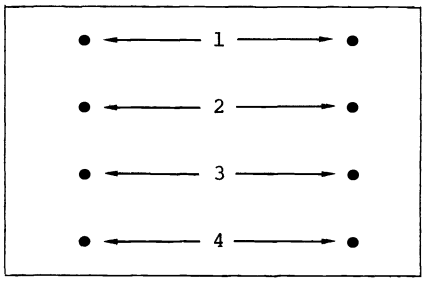


FIG. 1. *A 4-port.*

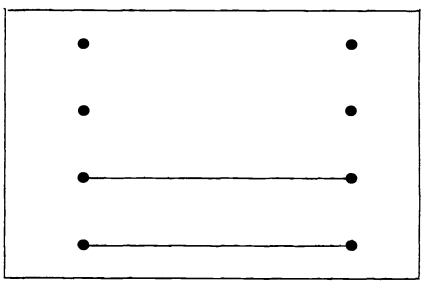


FIG. 2. *A shorted 4-port.*

2. Infinite dimensions and some notation. The shorted operator construction in infinite dimensions of Anderson [1], Anderson and Trapp [3] and Krein [16] is a generalization of the Schur complement to positive operators on a Hilbert space $\mathscr{H}$.

Let A be positive, i.e., $A = A^*$ and $(Ax, x) \geqq 0$. Let S be a closed subspace and partition A as before.

THEOREM [3]. *There is a unique operator $S(A)$ such that*

$$S(A) = \sup_X \left\{ X : 0 \leqq X, \begin{bmatrix} X & 0 \\ 0 & 0 \end{bmatrix} \leqq A \right\}.$$

The operator $S(A)$ agrees with the classical Schur complement if A_{22} is invertible. In fact

$$S(A) = \lim_{\varepsilon \downarrow 0} A_{11} - A_{12}(A_{22} + \varepsilon)^{-1} A_{21}.$$

If $\mathscr{H}$ is finite-dimensional, then the shorted operator construction agrees with the generalized Schur complement [2]. However, in infinite dimensions the shorted operator may exist (as a bounded operator) when the generalized Schur complement of Ando does not exist.

In the following sections $S(A)$ always refers to the shorted operator construction of Krein and Anderson.

3. The approximate equations for the shorted operator. In this section we give the following limiting equations

$$\lim_{n \to \infty} A_{11} c + A_{12} y_n = d,$$

$$\lim_{n \to \infty} A_{21} c + A_{22} y_n = 0,$$

$$(A_{22} y_n, y_n) \leqq M,$$

that uniquely define the shorted operator $S(A)c = d$.

The following proposition is a direct corollary of a result of Douglas [11]. Its proof may be found in [3].

PROPOSITION 1. *Let A be a positive operator where A is partitioned as*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Then there is a unique operator C such that

$$A_{21} = A_{22}^{1/2} C, \quad \textit{and}$$

$$\ker C^* \supseteq \ker A_{22}^{1/2}.$$

The following proposition gives a formula for the shorted operator due to Anderson and Trapp (see [3]).

PROPOSITION 2. *Let A be as above and let C be the unique operator satisfying $A_{21} = A_{22}^{1/2} C$ and* ker $C^* \supseteq$ ker $A_{22}^{1/2}$. *Then the shorted operator $S(A)$ is given by $S(A) = A_{11} - C^*C$.*

We are now in a position to prove our main result (Theorem 1 below) in a series of lemmas.

LEMMA 1. *Let A be positive. Partition A as*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Let $c \in \mathscr{H}$. Then there is a sequence (or net) $\{y_n\}$ and a real number M such that

$$A_{21}c + A_{22}y_n \to 0,$$

$$(A_{22}y_n, y_n) \leqq M, \quad \textit{and}$$

$$A_{11}c + A_{12}y_n \to S(A)c.$$

Proof. By Proposition 1, there is a unique operator C that satisfies both $A_{21} = A_{22}^{1/2}C$ and $\ker(C^*) \supseteq \ker(A_{22}^{1/2})$. From the latter condition we obtain $\text{range}(A_{22}^{1/2}) \subseteq \overline{\text{range}(C)}$. Thus we may choose $\{y_n\} \in S^{\perp}$ so that $\lim_{n\to\infty} A_{22}^{1/2}y_n = C(-c)$. It follows that

$$\lim_{n\to\infty} A_{21}c + A_{22}y_n = \lim_{n\to\infty} A_{22}^{1/2}Cc + A_{22}^{1/2}(A_{22}^{1/2}y_n)$$

$$= A_{22}^{1/2}Cc + A_{22}^{1/2}C(-c) = 0.$$

Similarly, after noting that $A_{12} = A_{21}^* = C^*A_{22}^{1/2}$, we compute $\lim_{n\to\infty} A_{11}c + A_{12}y_n = (A_{11} - C^*C)c = S(A)c$. Finally, $(A_{22}y_n, y_n) = \|A_{22}^{1/2}y_n\|^2$ which converges and hence is bounded. □

LEMMA 2. *Let A be positive and let S be a closed subspace. Partition A as above. Then for any sequence (or net) $\{y_n\}$, $d \in \mathscr{H}$, and $M \in \mathbb{R}$ satisfying*

$$A_{12}y_n \to d,$$

$$A_{22}y_n \to 0, \quad \textit{and}$$

$$(A_{22}y_n, y_n) \leqq M,$$

we have $A_{12}y_n \to 0$.

Proof. Since $(A_{22}y_n, y_n)$ is bounded, we may find a subsequence $\{y'_n\}$ of $\{y_n\}$ such that $A_{22}^{1/2}y'_n$ converges weakly to some $e \in S$. By the Hahn–Banach Theorem there is a sequence $\{z_n\}$ with $z_n \in$ convex hull $\{y'_i\}_{i=n}^{\infty}$ with

$$A_{12}z_n \to d,$$

$$A_{22}z_n \to 0,$$

$$A_{22}^{1/2}z_n \to e.$$

(It is not difficult to maintain the first two limits; remember z_n is in the convex hull of $\{y_i\}_{i=n}^{\infty}$.)

Since $x \mapsto (A_{22}x, x)$ is convex, it follows that $(A_{22}z_n, z_n) \leqq M$. From $A_{22}z_n = A_{22}^{1/2}(A_{22}^{1/2}z_n)$ it follows that $A_{22}^{1/2}e = 0$ and thus $A_{22}e = 0$.

Choose N_0 so large that $\|A_{22}^{1/2}z_n\| \leqq \|e\| + 1$ for all $n > N_0$. For any $\varepsilon > 0$ there is an $N > N_0$ such that for all $n \geqq N$

$$|(A_{22}^{1/2}z_n, A_{22}^{1/2}z_n - e)| \leqq \varepsilon(\|e\| + 1).$$

Consequently,

$$\lim_{n\to 0} (A_{22}^{1/2}z_n, A_{22}^{1/2}z_n - e) = 0;$$

however, since

$$(A_{22}^{1/2}z_n, e) = (z_n, A_{22}^{1/2}e) = 0,$$

it follows that $(A_{22}^{1/2}z_n, A_{22}^{1/2}z_n) \to 0$ and we conclude that $e = 0$.

Now for any $x \in \mathscr{H}$ and any $\lambda \in \mathbb{R}$ we have

$$\lim_{n\to\infty} \left(\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x \\ \lambda z_n \end{bmatrix}, \begin{bmatrix} x \\ \lambda z_n \end{bmatrix} \right) = (A_{11}x, x) + 2(d, x)\lambda.$$

If $d \neq 0$, then an appropriate choice of λ and x will make the above expression negative. The result follows. □

THEOREM 1. *Let $A \geqq 0$, $c \in \mathscr{H}$. Partition A as*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Let $\{y_n\} \in \mathscr{H}$, $M \in \mathbb{R}$, $d \in \mathscr{H}$ satisfy

$$A_{21}c + A_{22}y_n \to 0,$$

$$(A_{22}y_n, y_n) \leqq M,$$

$$A_{11}c + A_{12}y_n \to d;$$

then

$$A_{11}c + A_{12}y_n \to S(A)c.$$

Proof. Let z_n be the sequence guaranteed by Lemma 1. Then

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} c \\ z_n \end{bmatrix} \to \begin{bmatrix} S(A)c \\ 0 \end{bmatrix}$$

and

$$(A_{22}z_n, z_n) \leqq M.$$

Now let $w_n = y_n - z_n$. Since $x \mapsto (A_{22}x, x)$ is a (semi-) inner product, the sequence $(A_{22}w_n, w_n)$ is bounded. Now apply Lemma 2 and the result follows. □

COROLLARY 1. *Let A be positive and let $c \in \mathscr{H}$. Partition A as*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Let $y_n \in \mathscr{H}$, $M \in \mathbb{R}$, $d \in \mathscr{H}$ satisfy

$$\|A_{11}c + A_{12}y_n\| \leqq M,$$

$$\operatorname*{weak\,lim}_{n \to \infty} A_{21}c + A_{22}y_n = 0,$$

$$(A_{22}y_n, y_n) \leqq M;$$

then

$$\operatorname*{weak\,lim}_{n \to \infty} A_{11}c + A_{12}y_n = S(A)c.$$

Proof. Let $\{y'_n\}$ be any subsequence of $\{y_n\}$. Let $\{y''_n\}$ be a further subsequence of $\{y'_n\}$ such that

$$\operatorname*{weak\,lim}_{n \to \infty} A_{11}c + A_{12}y''_n = d.$$

Now let $\{z_n\}$ be a sequence with $z_n \in$ convex hull $\{y''_i\}_{i=n}^{\infty}$ and

$$A_{11}c + A_{12}z_n \to d,$$

$$A_{21}c + A_{22}z_n \to 0,$$

$$(A_{22}z_n, z_n) \leqq M.$$

(We can maintain boundedness of $(A_{22}z_n, z_n)$ because $x \mapsto (A_{22}x, x)$ is convex.) Now apply the previous theorem to conclude that $d = S(A)c$. □

4. An example. In this section we give an example of a positive operator A and a sequence $\{y_n\}$ which show that the boundedness of $(A_{22}y_n, y_n)$ is essential in Theorem 1 and its corollary.

Consider the following matrix:

$$A(\alpha,\beta)=\left[\begin{array}{c|cccc} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots \\ \hline \alpha & \beta & & & \\ \alpha^2 & & \beta^2 & & \\ \alpha^3 & & & \beta^3 & \\ \vdots & & & & \ddots \end{array}\right].$$

If $0 < \alpha, \beta < 1$, then $A(\alpha, \beta)$ is a bounded operator A: $l_2 \to l_2$. Let

$$\mathbf{x}=\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ \vdots \end{bmatrix}$$

be in l_2; then

$$(A(\alpha,\beta)\mathbf{x},\mathbf{x}) = 1 + 2\sum_{i=1}^{\infty}\alpha^i x_i + \sum_{i=1}^{\infty}\beta^i x_i^2.$$

However, by differential calculus, if $0 < \alpha, \beta < 1$, then

$$2\alpha^i x_i + \beta^i x_i^2 \geqq \frac{-\alpha^{2i}}{\beta^i}.$$

Therefore,

$$(A(\alpha,\beta)\mathbf{x},\mathbf{x}) \geqq 1 - \sum_{i=1}^{\infty}\left(\frac{\alpha^2}{\beta}\right)^i.$$

If $\alpha^2/\beta \leqq \frac{1}{2}$, it follows that $A(\alpha, \beta)$ is positive.

Now let $\alpha = \frac{1}{4}$ and $\beta = \frac{1}{8}$. By the above computation $A(\frac{1}{4}, \frac{1}{8})$ is positive. Setting

$$y_n=\begin{bmatrix} 0 \\ \vdots \\ 4^n \\ 0 \\ \vdots \end{bmatrix},$$

we have

$$A(\tfrac{1}{4},\tfrac{1}{8})\begin{bmatrix} 0 \\ y_n \end{bmatrix}=\begin{bmatrix} 1 \\ 0 \\ \vdots \\ \frac{1}{2}^n \\ 0 \\ \vdots \end{bmatrix},$$

while

$$\left(\begin{bmatrix} \beta & & \\ & \beta^2 & \\ & & \ddots \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 4^n \\ 0 \\ \vdots \end{bmatrix}, \begin{bmatrix} 0 \\ \vdots \\ 4^n \\ 0 \\ \vdots \end{bmatrix} \right) = 2^n.$$

Letting

$$A = A(\tfrac{1}{4}, \tfrac{1}{8}) = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

be the partition relative to the upper left (1×1) corner, we may restate the above calculations in the form

$$A_{11}0 + A_{12}y_n \to 1 \neq S(A)0,$$

$$A_{21}0 + A_{22}y_n \to 0,$$

$$(A_{22}y_n, y_n) = 4^n.$$

This shows that without the boundedness of $(A_{22}y_n, y_n)$ the conclusion of Theorem 1 may fail.

For completeness sake we point out that Proposition 2 enables us to compute

$$\begin{aligned} S(A) &= 1 - \sum_{i=1}^{\infty} (i^{-i}4^{-i/2})(8^{-i}4^{-i/2}) \\ &= 1 - \sum_{i=1}^{\infty} 2^{-i} \\ &= 0. \end{aligned}$$

5. Infinite networks. The rigorous treatment of infinite resistive networks does not have a long history. As shown by Flanders [13], [14], Zemanian [22]–[24] and others many of the classical theorems for finite networks fail for infinite networks. In this section we derive as a corollary to Theorem 1 an existence and uniqueness result for infinite networks similar to a result of Flanders [13]. Our formulation of Kirchhoff's laws follows [10].

Given an infinite connected directed graph with edge set B, node set N, we define its incidence matrix

$$E = \{e_{ij}\}_{i \in N, j \in B}$$

by

$$e_{ij} = \begin{cases} +1 & \text{if edge } j \text{ points to node } i, \\ -1 & \text{if edge } j \text{ points away from node } i, \\ 0 & \text{otherwise.} \end{cases}$$

Let 0 and 1 be two distinguished nodes. Assume that the graph is locally finite except perhaps at zero. Set

$$E' = \{e_{ij}\}_{i \in N \setminus \{0\}, j \in B},$$

the matrix obtained from E by deleting the row corresponding to node 0. Under the assumption of local finiteness, the rows of E' will have finite support. Thus E' induces a continuous operator

$$E'\colon l_2(B) \to l_2(N\setminus\{0\})$$

and also a continuous operator

$$E'\colon \mathbb{R}^B \to \mathbb{R}^{N\setminus\{0\}}.$$

Here $l_2(B)$ denotes the Hilbert space of all sequences $\{x_b\}_{b\in B}$ with $\sum_{b\in B}|x_b|^2 < \infty$. The notation $\mathbb{R}^B$ denotes unrestricted sequences $\{x_b\}_{b\in B}$ with the topology of coordinate-wise convergence.

Let $c \in l_2(N\setminus\{0\}) \subseteq \mathbb{R}^{N\setminus\{0\}}$ be the vector $c = \{c_i\}_{i\in N\setminus\{0\}}$ defined by $c_1 = 1$, $c_i = 0$, $i \neq 1$. Kirchhoff's current law (for a current source of one amp connected between 0 and 1) may be expressed as

$$E'x = c$$

where $x \in \mathbb{R}^B$ denotes the branch currents. Kirchhoff's voltage law becomes

$$v \in \text{range } (E'^*) = \ker (E')^{\perp}$$

where $v \in \mathbb{R}^B$ denotes the branch voltage drops [10]. (The fact that range $(E'^*) = \ker (E')^{\perp}$ follows from the Fredholm alternative in the l_2 case, the $\mathbb{R}^B$ case can be found in [13]. By orthogonal complement in $\mathbb{R}^B$ we mean the closure in $\mathbb{R}^B$ of the l_2 orthogonal complement.)

Let $r_i > 0$ denote the resistance in the ith branch. Assume $r_i < M$ for all i. Then the operator $Rx = v$, where $v_i = r_i x_i$ is continuous either on $l_2(B)$ or $\mathbb{R}^B$. Now the equations

$$\begin{aligned} E'x &= c, \\ Rx - E'^*w &= 0 \end{aligned} \tag{3}$$

formally express the laws of Kirchoff and Ohm.

THEOREM. *Assuming the hypothesis above, there is a unique $w \in \mathbb{R}^B$, $w \in$ range E'^*, such that there are $x_n \in l_2(B)$ with*

(1) $\lim Rx_n = w \in \ker (E')^{\perp}$,

(2) $E'x_n = c$,

(3) (Rx_n, x_n) *is bounded.*

Proof. Let $S = \ker (E')^{\perp}$. Now partition R as

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

with $R_{11}\colon S \to S$, etc. If the underlying graph is connected, then there is a $d \in S = (\ker E')^{\perp}$ with $E'd = c$. With respect to the above partition, the condition that $E'x = c$ may be rewritten as

$$x = \begin{bmatrix} d \\ y \end{bmatrix},$$

where $y \in S^{\perp} = \ker E'$. Thus (3) may be rewritten as

$$\begin{aligned} R_{11}d + R_{12}y_n &= w, \\ R_{21}c + R_{22}y_n &\to 0, \\ (R_{22}y_n, y_n) &\leqq M. \end{aligned} \tag{3'}$$

The existence of such a sequence $\{y_n\}$ is guaranteed by Lemma 1. By diagonalization choose a $z \in \mathbb{R}^B$ with $y_n \to z$ through a subsequence. Such a z exists, since for each $i \in B$, $r_i y_{n,i}^2 \leqq (Ry_n, y_n) \leqq M$. Existence now follows. For uniqueness, let $y_n \to z$ and $y'_n \to z'$ solve (3'), then $y_n - y'_n$ satisfies the conditions of Lemma 2. Following the proof of Lemma 2 we may replace $y_n - y'_n$ by sequences $\{z_n - z'_n\}$, with $z_n - z'_n \in$ convex hull $\{y_i - y'_i\}_{i=n}^{\infty}$, such that $(R_{22}(z_n - z'_n), (z_n - z'_n)) \to 0$, and such that the limits $z = \lim_{n\to\infty} z_n$ and $z' = \lim_{n\to\infty} z'_n$ are not disturbed. It now follows that $z = z'$. □

REFERENCES

[1] W. N. ANDERSON, *Shorted operators*, SIAM J. Appl. Math., 20 (1971), pp. 520–525.

[2] W. N. ANDERSON, T. D. MORLEY AND G. E. TRAPP, *Cascade addition and subtraction of matrices*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 609–626.

[3] W. N. ANDERSON AND G. E. TRAPP, *Shorted operators* II, SIAM J. Appl. Math., 28 (1975), pp. 60–71.

[4] T. ANDO, *Generalized Schur complements*, Linear Algebra Appl., 27 (1979), pp. 173–186.

[5] D. CARLSON, *Matrix decompositions involving the Schur complement*, SIAM J. Appl. Math., 28 (1975), pp. 577–587.

[6] ———, *What are Schur complements anyway?* Linear Algebra Appl., 59 (1984), pp. 189–193.

[7] D. CARLSON, E. HAYNSWORTH AND T. MARKAM, *A generalization of the Schur complement by means of the Moore–Penrose inverse*, SIAM J. Appl. Math., 26 (1974), pp. 254–259.

[8] D. CARLSON AND E. V. HAYNSWORTH, *Complementable and almost definite matrices*, Linear Algebra Appl., 52/53 (1983), pp. 157–176.

[9] R. W. COTTLE, *Manifestations of the Schur complement*, Linear Algebra Appl., 8 (1974), pp. 189–211.

[10] R. J. DUFFIN AND T. D. MORLEY, *Almost definite operators and electro-mechanical systems*, SIAM J. Appl. Math., 35 (1978), pp. 21–30.

[11] R. G. DOUGLAS, *On the majorization, factorization and range inclusion of operators in Hilbert spaces*, Proc. Amer. Math. Soc., 17 (1966), pp. 413–416.

[12] P. G. FILMORE AND J. P. WILLIAMS, *On operator ranges*, Adv. in Math., 7 (1971), pp. 254–281.

[13] H. FLANDERS, *Infinite networks:* I—*resistive networks*, IEEE Trans. Circuit Theory, CT-18 (1971), pp. 326–331.

[14] ———, *Infinite networks* II—*resistance in an infinite grid*, J. Math. Anal. Appl., 40 (1972), pp. 30–35.

[15] W. L. GREEN AND T. D. MORLEY, *Operator means, fixed points and the norm convergence of monotone approximants*, Math. Scand., to appear.

[16] M. G. KREIN, *The theory of self-adjoint extensions of semibounded Hermitian transformations and its applications* I *and* II, Mat. Sb. (N.S.) (Moscow), 20 (1947), pp. 431–495; 21 (1947), pp. 365–404.

[17] S. K. MITRA AND M. L. PURI, *Shorted operators—an extended concept and some applications*, Linear Algebra Appl., 47 (1982), pp. 57–59.

[18] T. D. MORLEY, *Several applications of the shorted operator*, in Constructive Approaches to Mathematical Models, C. Coffman and G. Fix, eds., Academic Press, New York, 1979.

[19] S. SESHU AND M. REED, *Linear Graphs and Electrical Networks*, Addison-Wesley, Reading, MA, 1961.

[20] G. E. TRAPP, *Hermitian semidefinite matrix means and related matrix inequalities—an introduction*, Linear and Multilinear Algebra, 16 (1984), pp. 113–123.

[21] L. WEINBERG, *Network Analysis and Synthesis*, McGraw-Hill, New York, 1962.

[22] A. H. ZEMANIAN, *Infinite electrical networks*, Proc. IEEE, 64 (1976), pp. 6–17.

[23] ———, *Limb analysis of infinite electrical networks*, Combin. Theory Ser. B, 24 (1978), pp. 76–93.

[24] ———, *Nonuniform semi-infinite grounded grids*, SIAM J. Math. Anal., 13 (1982), pp. 770–788.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

INEQUALITIES FOR THE TRACE OF MATRIX EXPONENTIALS*

DENNIS S. BERNSTEIN†

Abstract. Several inequalities involving the trace of matrix exponentials are derived. The Golden–Thompson inequality tr $e^{A+B} \leqq$ tr $e^A e^B$ for symmetric A and B is obtained as a special case along with the new inequality tr $e^A e^{A^T} \leqq$ tr e^{A+A^T} for nonnormal A.

Key words. matrix exponential, inequality, trace

AMS(MOS) subject classification. 15

1. Introduction. For $n \times n$ real symmetric matrices A and B, the Golden–Thompson inequality [1]–[5] states that

$$\text{tr } e^{A+B} \leqq \text{tr } e^A e^B. \tag{1.1}$$

Reference [5] generalizes (1.1) to allow arbitrary spectral functions in place of the trace operator and provides an overview of applications in which these inequalities arise.

In contrast to (1.1), problems in linear-quadratic optimal feedback control [6] typically involve a performance functional J of the form

$$J = \text{tr} \int_0^\infty e^{At} V e^{A^T t} R \, dt, \tag{1.2}$$

where V and R denote noise intensity and performance weighting matrices, respectively, and A denotes the linear system dynamics matrix. The form of (1.2) thus suggests inequalities of the form (1.1) involving A and A^T, where A is nonnormal, in place of symmetric A and B. Such inequalities are motivated by robust sampled-data control-design problems which require performance bounds for uncertain system models.

The main result of the present note is the inequality

$$\text{tr } e^A e^{A^T} \leqq \text{tr } e^{A+A^T}. \tag{1.3}$$

Rather surprisingly, the sign of the inequality (1.3) is opposite to the sign of (1.1). To understand why this is the case, we derive a series of inequalities which, upon appropriate specialization, yield both (1.1) and (1.3).

2. Inequalities. The following lemma is required. (Let C^T denote the transpose of a matrix C.)

LEMMA 2.1. *If $C \in R^{n \times n}$ and r is a positive integer, then*

$$\text{tr } C^{2r} \leqq \text{tr } C^r C^{rT} \leqq \text{tr } (CC^T)^r. \tag{2.1}$$

Proof. The first inequality follows from tr $(C^r - C^{rT})(C^{rT} - C^r) \geqq 0$, while the second follows from a result of K. Fan (see [4, pp. 234, 516]). □

THEOREM 2.1. *If $A, B \in R^{n \times n}$, then*

$$\begin{aligned} \text{tr } e^{A+B} &\leqq \text{tr } e^{(A+B)/2} e^{(A+B)^T/2} \leqq \text{tr } e^{(A+A^T+B+B^T)/2} \\ &\leqq \text{tr } e^{(A+A^T)/2} e^{(B+B^T)/2} \leqq \tfrac{1}{2} \text{tr } (e^{A+A^T} + e^{B+B^T}), \end{aligned} \tag{2.2}$$

$$\left.\begin{matrix} \text{tr } e^A e^B \\ \tfrac{1}{2} \text{tr } (e^{2A} + e^{2B}) \end{matrix}\right\} \leqq \tfrac{1}{2} \text{tr } (e^A e^{A^T} + e^B e^{B^T}) \leqq \tfrac{1}{2} \text{tr } (e^{A+A^T} + e^{B+B^T}). \tag{2.3}$$

* Received by the editors January 21, 1987; accepted for publication May 11, 1987. This research was supported in part by the Air Force Office of Scientific Research under contract F49620-86-C-0002.

† Harris Corporation, Melbourne, Florida 32902.

Proof. Defining $C = e^{A/2r}e^{B/2r}$, (2.1) becomes

$$\operatorname{tr}(e^{A/2r}e^{B/2r})^{2r} \leqq \operatorname{tr}(e^{A/2r}e^{B/2r})^{r}(e^{B^T/2r}e^{A^T/2r})^{r} \leqq \operatorname{tr}(e^{A/2r}e^{B/2r}e^{B^T/2r}e^{A^T/2r})^{r}.$$

Letting $r \to \infty$, the exponential product formula [5, p. 60] and its generalization [7, p. 97] yield the first two inequalities of (2.2). The third inequality of (2.2) follows from Corollary 3 of [5] while the fourth inequality of (2.2) follows from

$$0 \leqq \operatorname{tr}[e^{(A+A^T)/2} - e^{(B+B^T)/2}]^2.$$

To prove (2.3) note that the upper leftmost inequality follows from $0 \leqq \operatorname{tr}(e^A - e^B)(e^A - e^B)^T$. The remaining inequalities in (2.3) follow from $\operatorname{tr} e^{2A} \leqq \operatorname{tr} e^A e^{A^T} \leqq \operatorname{tr} e^{A+A^T}$, which is a consequence of (2.2) with $B = A$. □

COROLLARY 2.1. *If* $A \in R^{n\times n}$, *then*

$$\operatorname{tr} e^{2A} \leqq \operatorname{tr} e^A e^{A^T} \leqq \operatorname{tr} e^{A+A^T} \leqq \frac{n}{2} + \frac{1}{2}\operatorname{tr} e^{2(A+A^T)}, \tag{2.4}$$

$$\operatorname{tr} e^{2A} \leqq \frac{n}{2} + \frac{1}{2}\operatorname{tr} e^{2A}e^{2A^T} \leqq \frac{n}{2} + \frac{1}{2}\operatorname{tr} e^{2(A+A^T)}. \tag{2.5}$$

If $A, B \in R^{n\times n}$ *are symmetric, then*

$$\operatorname{tr} e^{A+B} \leqq \operatorname{tr} e^A e^B \leqq \tfrac{1}{2}\operatorname{tr}(e^{2A} + e^{2B}). \tag{2.6}$$

Proof. The first two inequalities of (2.4) follow from the first two inequalities of (2.2) with $B = A$. The last inequality of (2.4) follows from the last inequality of (2.2) with $B = 0$ and A replaced by $2A$. Inequalities (2.5) follow from (2.3) with $B = 0$ and A replaced by $2A$ while ignoring the lower leftmost term in (2.3). Finally, (2.6) follows from (2.2). □

Remark. The second inequality in (2.4) and the first inequality in (2.6) correspond to (1.3) and (1.1), respectively.

3. Additional inequalities. The question immediately arises as to whether any additional inequalities involving the expressions appearing in (2.4) and (2.5) are true. Note that $\operatorname{tr} e^A e^B$ in (2.3) cannot be merged with (2.2) because of the sign reversal between (1.1) and (1.3). It can readily be seen that the only remaining possibilities are

$$\operatorname{tr} e^{(A+A^T)/2}e^{(B+B^T)/2} \stackrel{?}{\leqq} \tfrac{1}{2}\operatorname{tr}(e^A e^{A^T} + e^B e^{B^T}), \tag{3.1}$$

$$\operatorname{tr} e^{(A+A^T)/2}e^{(B+B^T)/2} \stackrel{?}{\leqq} \tfrac{1}{2}\operatorname{tr}(e^{2A} + e^{2B}), \tag{3.2}$$

$$\operatorname{tr} e^A e^B \stackrel{?}{\leqq} \tfrac{1}{2}\operatorname{tr}(e^{2A} + e^{2B}). \tag{3.3}$$

By randomly generating A and B, (3.1) was shown to be false. Since (3.2) implies (3.1), (3.2) must also be false. Furthermore, in the case $B^T = -B$, inequality (3.1), which becomes

$$\operatorname{tr} e^{(A+A^T)/2} \stackrel{?}{\leqq} \frac{n}{2} + \frac{1}{2}\operatorname{tr} e^A e^{A^T}, \tag{3.4}$$

was also shown to be false. Hence (2.4) and (2.5) cannot be merged. Finally, the remaining inequality (3.3) was also shown to be false even when $B = 0$.

Remark. The results of this paper can be generalized to the case in which A and B are complex matrices. Generalization to arbitrary spectral functions [5] remains an area for further research.

Acknowledgment. I wish to thank Scott W. Greeley for carrying out numerical calculations which suggested the results of this paper.

REFERENCES

[1] S. GOLDEN, *Lower bounds for the Helmholtz function*, Phys. Rev., 137 (1965), pp. B1127–B1128.

[2] C. J. THOMPSON, *Inequality with applications in statistical mechanics*, J. Math. Phys., 6 (1965), pp. 1812–1813.

[3] A. LENARD, *Generalization of the Golden–Thompson Inequality* $\operatorname{Tr}(e^A e^B) \geqq \operatorname{Tr} e^{A+B}$, Indiana Univ. Math. J., 21 (1971), pp. 457–467.

[4] A. W. MARSHALL AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[5] J. E. COHEN, S. FRIEDLAND, T. KATO, AND F. P. KELLY, *Eigenvalue inequalities for products of matrix exponentials*, Linear Algebra Appl., 45 (1982), pp. 55–95.

[6] K. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, John Wiley, New York, 1972.

[7] V. S. VARADARAJAN, *Lie Groups, Lie Algebras, and Their Representations*, Springer-Verlag, Berlin, New York, Heidelberg, 1984.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

003

COMPLETION OF TOEPLITZ PARTIAL CONTRACTIONS*

CHARLES R. JOHNSON† AND LEIBA RODMAN‡

Abstract. Those patterns for the specified entries of a partial Toeplitz matrix (whose specified entries occur on consecutive diagonals) are characterized, which ensure that a Toeplitz partial contraction may be completed to a Toeplitz contraction. The answer is rather different from that of the corresponding question without the Toeplitz condition.

Key words. contraction, matrix completion, partial matrix, Toeplitz matrix, specified entries

AMS(MOS) subject classifications. 47A20, 15A60, 15A57, 47A30

1. Introduction and the main result. An m-by-n complex matrix $A = (a_{ij})$ is called a *Toeplitz matrix* if

$$a_{ij} = a_{j-i}$$

for some sequence of $m + n - 1$ complex numbers

$$a_{-(m-1)}, \cdots, a_0, \cdots, a_{n-1},$$

i.e., a Toeplitz matrix is constant along upper left-to-lower right diagonals. An m-by-n complex matrix B is called a *contraction* if $I - BB^*$ is positive semidefinite. Equivalently, each eigenvalue of BB^* (or singular value of B) is $\leqq 1$. The m-by-n matrix A is called a *Toeplitz contraction* if it is both a Toeplitz matrix and a contraction.

By a *partial matrix* we mean an m-by-n array A, some of whose entries are *specified* complex numbers, and whose remaining entries are *unspecified* (i.e., free variables whose values are to be chosen from the complex numbers). A *completion* of a partial matrix is simply a particular specification of the unspecified entries resulting in a conventional matrix. For example,

$$\begin{bmatrix} -i & 2 & 1 \\ 0 & -i & 2 \end{bmatrix}$$

is a completion of the 2-by-3 partial matrix

$$\begin{bmatrix} -i & 2 & ? \\ ? & -i & 2 \end{bmatrix}.$$

We adopt the convention of denoting unspecified entries by ?'s. When just the placement of specified entries is to be indicated, we denote them with X's as in

$$\begin{bmatrix} X & X & ? \\ ? & X & X \end{bmatrix}.$$

Partial matrices have been discussed in [1]–[5].

A *partial Toeplitz matrix* is simply an m-by-n partial matrix $A = (a_{ij})$ such that if a_{ij} is specified, then a_{kl} is specified and equal to a_{ij} for each pair k, l with $k - l = i - j$,

* Received by the editors September 2, 1986; accepted for publication (in revised form) June 1, 1987.

† Mathematics Department, College of William and Mary, Williamsburg, Virginia 23185. The work of this author was supported in part by National Science Foundation grant DMS-8713762 and by Office of Naval Research contract N00014-86-K-0012.

‡ Mathematics Department, Arizona State University, Tempe, Arizona 85287, and School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. Present address, Mathematics Department, College of William and Mary, Williamsburg, Virginia 23185. The work of this author was partially supported by National Science Foundation grant DMS-8501794 and by a Mini Grant from the College of Liberal Arts and Sciences, Arizona State University.

i.e., A is Toeplitz to the extent it is specified. The first partial matrix mentioned above is a partial Toeplitz matrix. A *partial contraction* is a partial matrix each of whose rectangular submatrices consisting entirely of specified entries (i.e., specified submatrices) is itself a contraction. If a partial matrix is to have a completion which is a contraction, it must be a partial contraction. A *Toeplitz partial contraction* is a partial matrix that is both a partial contraction and partial Toeplitz.

We are interested in which Toeplitz partial contraction have completions that are Toeplitz contractions. In [5] we studied the question of which partial contractions have completions that are contractions. Here we address the Toeplitz version of this question in a similar qualitative manner, a question suggested to us by C. Foias, who pointed out that problems of this type arise in the analysis of input-output response in layered media.

We shall often be interested in the pattern P for the specified entries of (a class of) partial matrices. In the case of partial Toeplitz matrices, this would be an indication of which diagonals are specified and which are unspecified. We adopt the convention of numbering these diagonals so that the i, j cell lies on diagonal $j - i$, i.e., an m-by-n array has $m + n - 1$ diagonals numbered $-(m - 1), \cdots, 0, \cdots, (n - 1)$ from the lower left to upper right.

We address the question of which patterns for the specified entries of a Toeplitz partial contraction guarantee that it may be completed to a Toeplitz contraction. We give a complete answer in the case in which the specified diagonals form exactly a consecutive sequence of diagonals. Note that the number d of specified diagonals is always $\leqq m + n - 1$ for an m-by-n partial Toeplitz matrix A. Our result is the following.

THEOREM 1. *Let P be an m-by-n partial Toeplitz pattern whose d specified diagonals form a consecutive sequence. Then every partial Toeplitz contraction with specified entries of pattern P has a Toeplitz contraction completion if and only if at least one of the following conditions is satisfied*:

(i) $m = 1$ *or* $n = 1$,
(ii) $d \leqq 1$,
(iii) $d \geqq m + n - 2$, *or*
(iv) $m = n$ *and* $d = 2n - 3$.

According to the theorem, the only partial Toeplitz patterns (up to transposition) that ensure completion of a partial Toeplitz contraction to a Toeplitz contraction are the following:

(i) a single row,
(ii)

$$\begin{bmatrix} X & & & & & \\ & X & & & & \\ & & X & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & X \end{bmatrix},$$

(iii)

$$\begin{bmatrix} X & \cdot & \cdot & \cdot & \cdot & X & ? \\ \cdot & \cdot & & & & & X \\ \cdot & & & & & & \cdot \\ \cdot & & \cdot & & & & \cdot \\ \cdot & & & & & & \cdot \\ X & \cdot & \cdot & X & \cdot & \cdot & X \end{bmatrix}, \quad \text{and}$$

(iv)(a)

$$\begin{bmatrix} X & \cdots & X & ? \\ \cdot & & \cdot & X \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ X & & \cdot & \cdot \\ & & \cdot & \cdot \\ ? & X & X & X \end{bmatrix} \quad \text{or}$$

(iv)(b)

$$\begin{bmatrix} X & \cdot & \cdot & \cdot & X & ? & ? \\ \cdot & & & & & X & ? \\ \cdot & & & & & & X \\ \cdot & & & & & & \cdot \\ & & & & & & \cdot \\ & & & & & & \cdot \\ X & \cdot & \cdot & \cdot & & & X \end{bmatrix} \quad \text{in the square case.}$$

It should be noted that, although the completable patterns are on the whole more restrictive in the Toeplitz case, the answer is quite different from the general contraction case [5]. The pattern (iv)(a) is *not* a completable pattern in the general contraction case, while many patterns allowed in the general case [5] are excluded in the Toeplitz version.

It should be noted also that, as follows from Theorem 1, not every Toeplitz subpattern (with obvious definition of this notion) of a completable Toeplitz pattern is completable. This is in contrast with the case of general (non-Toeplitz) contraction completions (see [5]).

2. Proof of Theorem 1 (sufficiency). The general strategy of proof is similar to the characterization of completable patterns in other settings [3]–[5]: verify the completability of the identified patterns (in this case this is relatively straightforward) and then present a class of counterexamples that show that no other patterns can be augmented without a contradiction (in this case this part is relatively more intricate than in [5]).

We first note that each of the patterns (i)–(iv) permits completion of a Toeplitz partial contraction to a partial contraction. In cases (i) and (ii) choosing the unspecified entries to be 0 is easily seen to produce a contraction if the partial matrix is a partial contraction. For (iii) and (iv) we recall the basic result (see [1], [5]) that in the general (not necessarily Toeplitz) case the block pattern

$$\begin{bmatrix} X & ? \\ X & X \end{bmatrix}$$

for a partial contraction always permits a contraction completion. Case (iii) follows directly from this fact; since only one entry is unspecified, the Toeplitz restriction on the completion is irrelevant. Case (iv) is slightly more subtle. In case (iv)(a) note that the upper right ? completes two maximal submatrices: the top $n-1$ rows and the last $n-1$ columns, but these two are essentially the same because of the Toeplitz restriction. Thus, specification of the ? using the above fact for one of these will necessarily satisfy the other and result in a Toeplitz partial contraction with only a ? in the lower left. An appeal to (iii) then completes the analysis of (iv)(a). Case (iv)(b) is similar. Again the two maximal submatrices completed by the specification of diagonal number $n-2$ are the same due to the Toeplitz condition and this diagonal may be specified (using the same fact again

applied to appropriate submatrices) to produce a Toeplitz partial contraction of type (iii), thus completing case (iv)(b) and the sufficiency of cases (i)–(iv).

The proof that no further patterns always allow completion to a Toeplitz contraction is based upon what might be called the *principle of incompatible specification.* (Though not formalized, this has been the guiding notion in prior work [3]–[5].) Note that in the sufficiency of cases (iii) and (iv) above, although specification of an unspecified diagonal sometimes completed more than one maximal submatrix, these submatrices were essentially the same; and, thus, the conditions upon the entry (diagonal) to be specified were compatible. In all other patterns, specification of a "next" diagonal completes more than one maximal and essentially different submatrix. It turns out that data for the specified entries may always be exhibited in such cases so that the various conditions placed upon the next diagonal are incompatible. In fact, this may be done so that two of the different completed submatrices uniquely determine the new entry to be two different values. This is the incompatible specification principle.

3. Auxiliary results. We shall develop some auxiliary results of computational character in order to put the incompatible specification principle to work.

LEMMA 2. *Let $G(x, y)$ be a q-by-r matrix (where $1 \leqq q \leqq r$) all of whose elements, with the exception of one diagonal consisting of q elements, are equal to y, and these exceptional entries are equal to x (here $x, y \in \mathbf{C}$). Then $G(x, y)$ is a contraction if and only if the following conditions hold*:

(1) *For $q \geqq 2$, either*

$$|x-y|<1 \quad \text{and} \quad (r-2)|y|^2+x\bar{y}+y\bar{x} \leqq q^{-1}(1-|x-y|^2)$$

or

$$|x-y|=1 \quad \text{and} \quad (r-2)|y|^2+x\bar{y}+y\bar{x}=0.$$

(2) *For $q = 1$,*

$$(r-1)|y|^2+|x|^2 \leqq 1.$$

In particular, $G(x, x)$ is a contraction if and only if $|x| \leqq (qr)^{-1/2}$.

Proof. Leaving aside the trivial case $q = 1$, assume that $q \geqq 2$. A calculation shows that the q-by-q matrix $I - G(x, y)G(x, y)^*$ is equal to

$$(1-|x-y|^2)I-((r-2)|y|^2+x\bar{y}+y\bar{x})e_q^* e_q \tag{1}$$

where e_q is the 1-by-q row all of whose entries are 1. Since $q \geqq 2$, in order that (1) be positive semidefinite it is necessary that $|x - y| \leqq 1$. The case $|x - y| = 1$ being evident, assume that $|x - y| < 1$. Then (1) is positive semidefinite if and only if $I - we_q^* e_q$ is such, where

$$w=(1-|x-y|^2)^{-1}((r-2)|y|^2+x\bar{y}+y\bar{x}).$$

Now

$$\langle (I-we_q^* e_q)x, x\rangle = \langle x, x\rangle - w|\langle e_q^*, x\rangle|^2 \tag{2}$$

for every $x \in \mathbf{C}^q$. (Here $\langle \cdot, \cdot \rangle$ stands for the standard inner product on $\mathbf{C}^q$.) It follows that (2) is nonnegative for every $x \in \mathbf{C}^q$ if and only if

$$1-w\langle e_q^*, e_q^*\rangle \geqq 0$$

which means that $w \leqq q^{-1}$.

LEMMA 3. *Let $F_{q,r}$ be a $q \times r$ matrix ($q \geqq r \geqq 1$), whose every entry, with the exception of one unspecified entry in the upper right-hand corner, is equal to*

$[(q-1)r]^{-1/2}$. *Then there is a unique contraction completion of* $F_{q,r}$, *and it is obtained by specifying the upper right-hand corner to be*

$$w_0 = -(r-1)[(q-1)r]^{-1/2}. \tag{3}$$

Proof. It follows from Lemma 2 that $F_{q,r}$ is a partial Toeplitz contraction. Let $F(x)$ be the q-by-r matrix obtained from $F_{q,r}$ by specifying the upper right-hand corner to be $x \in \mathbf{C}$. Then, denoting $w = [(q-1)r]^{-1/2}$, we have

$$I - F(x)F(x)^* = \begin{bmatrix} 1-(r-1)|w|^2-|x|^2 & [-(r-1)|w|^2 - x\bar{w}]e_{q-1} \\ [-(r-1)|w|^2 - \bar{x}w]e_{q-1}^* & V \end{bmatrix}$$

where $e_{q-1} = [1 \cdots 1]$ is 1-by-$(q-1)$ row and

$$V = I - r|w|^2 e_{q-1}^* e_{q-1} = I - (q-1)^{-1} e_{q-1}^* e_{q-1}$$

is $(q-1)$-by-$(q-1)$ matrix. One verifies that e_{q-1}^* belongs to the kernel of V. Let U be a $(q-1)$-by-$(q-1)$ unitary matrix whose first column is $e_{q-1}^*/\|e_{q-1}^*\|$. Then the 2-by-2 upper left-hand corner in

$$\begin{bmatrix} 1 & 0 \\ 0 & U^* \end{bmatrix}(I - F(x)F(x)^*)\begin{bmatrix} 1 & 0 \\ 0 & U \end{bmatrix}$$

is

$$\begin{bmatrix} 1-(r-1)|w|^2-|x|^2 & (q-1)^{1/2}[-(r-1)|w|^2 - x\bar{w}] \\ (q-1)^{1/2}[-(r-1)|w|^2 - \bar{x}w] & 0 \end{bmatrix}. \tag{4}$$

The matrix (4) is positive semidefinite only if

$$-(r-1)|w|^2 - x\bar{w} = 0,$$

or

$$x = -(r-1)w.$$

So, if there exists a contraction completion of $F_{q,r}$, it must be specified by (3). But the existence of a contraction completion of $F_{q,r}$ follows from [1], [5], and the lemma is proved.

Using Lemma 2 and calculations similar to that in the proof of Lemma 3, we obtain the following statement.

LEMMA 4. *Let* $H(x, y)$ *be a q-by-r partial matrix (where* $2 \leqq q \leqq r-1$*) all of whose entries except for the upper right-hand corner are specified. The specified entries are equal to y, except for one diagonal consisting of q elements which does not intersect the last column in* $H(x, y)$*; the entries on this exceptional diagonal are equal to x. Here x and y are complex numbers such that*

$$|x-y| < 1, \qquad y \neq 0,$$
$$(r-2)|y|^2 + x\bar{y} + \bar{x}y = (q-1)^{-1}(1-|x-y|^2),$$
$$(r-3)|y|^2 + x\bar{y} + \bar{x}y \leqq q^{-1}(1-|x-y|^2).$$

Then there is a unique number z such that by specifying the upper right-hand corner in $H(x, y)$ *to be z, one obtains a contraction. This number is given by the formula*

$$z = [-(r-3)|y|^2 - x\bar{y} - \bar{x}y]\bar{y}^{-1}.$$

We remark that the statements in Lemmas 3 and 4 concerning the values of w_0 and z, respectively, can also be obtained from a general formula for contraction completions [1, Thm. 1.2].

4. Proof of Theorem 1 (necessity). Let P be an m-by-n pattern of specified entries that is not covered by (i)–(iv) of Theorem 1. We shall show that not every partial Toeplitz contraction of pattern P admits a Toeplitz contraction completion. By passing to a subpattern, we can assume that $2 \leqq m < n$ and either

$$P = \begin{bmatrix} & ? \\ ? & \end{bmatrix} \quad \text{or} \tag{5}$$

$$P = \begin{bmatrix} ?? \\ ? \end{bmatrix}. \tag{6}$$

Consider first the case P as given by (6), and let $3 \leqq m \leqq n-2$. Let K be an m-by-n partial matrix with pattern P with the following properties:

(1) The specified entries in K, with the exception of one diagonal consisting of m elements which does not intersect the two last columns of K, are equal to

$$y = ((m-1)(m-2) + \tfrac{1}{4}(n+1-m)^2)^{-1/2}.$$

(2) The specified entries on the exceptional diagonal in K are equal to

$$x = \frac{m-n+1}{2} y.$$

One verifies that

$$|x-y| < 1,$$
$$(n-2)y^2 + 2xy = (m-2)^{-1}(1-|x-y|^2),$$
$$(n-3)y^2 + 2xy = (m-1)^{-1}(1-|x-y|^2),$$
$$(n-4)y^2 + 2xy < m^{-1}(1-|x-y|^2).$$

So by Lemma 4, K is a partial Toeplitz contraction which does not admit a Toeplitz contraction completion.

Assume now that $m = n-1 \geqq 3$ (and P is given by (6)). Let K be a partial Toeplitz matrix with pattern P each entry of which is $((n-2)(n-1))^{-1/2}$. By Lemma 2, K is a partial contraction. If K were to admit a Toeplitz contraction completion, then by Lemma 3 the only possibility in the $(1, n-1)$, hence also in the $(2, n)$ entry, would be

$$-(n-2)((n-2)(n-1))^{-1/2}.$$

However, this is impossible, because one easily checks that the n-dimensional row

$$[ww \cdots ww_0],$$

where $w = ((n-2)(n-1))^{-1/2}$, $w_0 = -(n-2)w$, is not a contraction for $n > 1$.

It remains to consider (still assuming P has the form (6)) the case $m = 2$, $n > 2$. If $n = 3$, then we are done by letting the specified entries be $\sqrt{2}/2$. If $n \geqq 4$, then put

$$y = \left(n-2+\frac{1}{4}(n-5)^2\right)^{-1/2}, \qquad x = -\frac{n-5}{2} y,$$

and verify that

$$|x-y|<1,$$
$$(n-4)y^2+2xy=\tfrac{1}{2}(1-|x-y|^2),$$
$$(n-2)y^2+x^2=1.$$

Let K be the partial Toeplitz matrix with pattern P, all of whose specified entries are y, with the exception of (1, 1) and (2, 2) entries which are x. By Lemma 2, K is a partial contraction. Arguing as in the proof of Lemma 4, we see that the only way K can be completed to a Toeplitz contraction is by putting

$$z=-(n-4)y-2x$$

in the $(1, n-1)$ position. However, the $1 \times n$ matrix

$$[yxy \cdots yz]$$

is not a contraction, so K does not admit a Toeplitz contraction completion.

Consider now the case when P is given by (5). If $3 \leqq m \leqq n-2$, then we are done by arguing as in the case when P has the form (6). Assume that $m = n-1 \geqq 2$. Let K be the partial matrix with the pattern P all of whose specified entries are $((n-2)(n-1))^{-1/2}$. Then K is a partial Toeplitz contraction by Lemma 2. By the principle of incompatible specifications, and using Lemma 3, K is not completable to a Toeplitz contraction. Finally assume that $m = 2$, $n \geqq 4$. Let

$$y=\left(n-2+\frac{1}{4}(n-5)^2\right)^{-1/2}, \qquad x=-\frac{n-5}{2}y,$$

and let K be the partial matrix with pattern P all of whose specified entries are y except for the (1, 2) and (2, 3) entries which are x. As in the case when P was given by (6), one verifies that K is a partial Toeplitz contraction which is not completable to a Toeplitz contraction.

Theorem 1 is now proved completely.

5. Partial Toeplitz contractions with scattered diagonals. In this section we consider briefly the Toeplitz completion problem for the case of scattered diagonals.

Given an $m \times n$ matrix A with entries a_{ij} $(1 \leqq i \leqq m;\ 1 \leqq j \leqq n)$, the diagonals of A that are parallel to the main diagonal will be numbered from $-(m-1)$ to $n-1$, starting in the lower left-hand corner and ending in the upper right-hand corner. Thus, the diagonal $\{a_{ij} | -i+j=d\}$, where d is fixed, has number d. Let

$$T(m, n; d_1, d_2, \cdots, d_r)$$

be the set of all $m \times n$ partial Toeplitz contractions whose specified diagonals have numbers $d_1, \cdots, d_r$ (it will be assumed that $d_1 < d_2 < \cdots < d_r$). The set

$$T(m, n; d_1, d_2, \cdots, d_r)$$

will be called *completable* if every matrix from $T(m, n; d_1, d_2, \cdots, d_r)$ admits a Toeplitz contraction completion.

Some completable sets (apart from those described in Theorem 1) are given below.

As the cases when $r = 1$ or $\min(m, n) = 1$ were covered in Theorem 1, we assume in the following theorem that $r \geqq 2$ and both m, n are greater than one.

THEOREM 5. *Assume that $d_1, \cdots, d_r$ form an arithmetic progression, i.e., $d_j = d_1 + k(j-1)$, $j = 2, \cdots, r$, for some $k > 0$ (independent of j), and assume that $d_1 =$*

$-\varphi k$ for some integer φ. Then $T(m, n; d_1, \cdots, d_r)$ is completable if and only if at least one of the following conditions is satisfied (here the integer ψ is defined by $d_r = \psi k$):

(1) $1 + (\varphi + 1)k > m$ *and* $1 + (\psi + 2)k > n$;

(2) $1 + (\varphi + 2)k > m$ *and* $1 + (\psi + 1)k > n$;

(3) $1 + (\psi + 2)k > m \geqq 1 + (\psi + 1)k$, $1 + (\psi + 2)k > n \geqq 1 + (\psi + 1)k$ *and* $\psi = 4$;

(4) $1 + (\psi + 1)k > m$, $1 + (\psi + 3)k > n \geqq 1 + (\psi + 2)k$ *and* $\psi = 4$;

(5) $1 + (\psi + 1)k > n$, $1 + (\psi + 3)k > m \geqq 1 + (\psi + 2)k$ *and* $\psi = 4$.

Proof. Assume at least one of the conditions (1)–(5) is satisfied. Given a partial Toeplitz contraction $A \in T(m, n; d_1, \cdots, d_r)$, let $\hat{A}$ be the matrix formed by the entries (specified or not) of A in the positions $(sk + 1, tk + 1)$, where s, t are integers. By Theorem 1 the partial Toeplitz contraction $\hat{A}$ is completable to a Toeplitz contraction. It is easy to see that by using this Toeplitz contraction completion in the entries $(sk + 1, tk + 1)$ of A (where s, t are integers), and by putting zeros in all the remaining entries of A, a Toeplitz contraction completion of A is produced.

Conversely, if none of the conditions (1)–(5) is satisfied, then by Theorem 1 there exists a partial Toeplitz contraction $A \in T(m, n; d_1, \cdots, d_r)$ such that $\hat{A}$ (constructed as above) is not completable to a Toeplitz contraction. Then, obviously, A cannot be completed to a Toeplitz contraction as well. □

One can describe, using the same idea as in the proof of Theorem 5, all completable sets $T(m, n; d_1, \cdots, d_r)$ where d_j's form an arithmetic progression but d_1 is not necessarily an integer multiple of k. However, this description is messy and hence will not be presented here.

We were not able to describe the completable sets $T(m, n; d_1, \cdots, d_r)$ in general. We propose the following conjecture.

CONJECTURE 6. Assuming that $r > 1$ and both m and n are larger than one, all completable sets $T(m, n; d_1, \cdots, d_r)$, possibly with very few exceptions, are those described in Theorem 5.

It should be noted that the "one-step extension" approach (see [2], [4], [6]) does not work for the Toeplitz contraction extension problem, as can be seen already from Theorem 1.

6. Final remarks. We conclude this paper with several observations concerning the main result (Theorem 1).

First, Theorem 1 is true also if all partial Toeplitz contractions involved, as well as their Toeplitz contraction completions, are assumed to have real entries. This is proved in exactly the same way as Theorem 1.

Another version of Theorem 1 can be obtained by replacing contractions with strict contractions (an m-by-n matrix A is called a strict contraction if $I - AA^*$ is positive definite). In this case the sufficiency part is proved as in the proof of Theorem 1. For the necessity part one has to modify slightly the arguments given in § 4. We omit the details.

Finally, let H be a Hilbert space. An m-by-n matrix $A = (A_{ij})$, where $A_{ij} : H \to H$ are linear bounded operators, is called *block operator Toeplitz* if $A_{ij} = A_{j-i}$ for some operators $A_{-(m-1)}, \cdots, A_0, \cdots, A_{n-1}$. Now the notions of a partial block operator Toeplitz contraction and of a block operator Toeplitz completion can be obviously defined, and Theorem 1 is true, together with its proof, in this framework.

REFERENCES

[1] 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.

[2] H. DYM AND I. GOHBERG, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36 (1981), pp. 1–24.

[3] R. GRONE, C. R. JOHNSON, E. SÁ, AND H. WOLKOWICZ, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.

[4] C. R. JOHNSON AND L. RODMAN, *Inertia possibilities for completions of partial Hermitian matrices*, Linear and Multilinear Algebra, 16 (1984), pp. 179–195.

[5] ———, *Completion of partial matrices to contractions*, J. Funct. Analysis, 69 (1986), pp. 260–267.

[6] R. L. ELLIS, I. GOHBERG, AND D. C. LAY, *Invertible self adjoint extensions of band matrices and their entropy*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 483–500.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

004

SEMI-ITERATIVE AND ITERATIVE METHODS FOR SINGULAR M-MATRICES*

G. P. BARKER† AND S.-J. YANG‡

Abstract. This paper provides a theoretical basis for establishing the convergence of semi-iterative and iterative techniques, especially Jacobi and Gauss–Seidel techniques, for computing nontrivial solutions of $Ax = 0$ where A is a singular M-matrix. These results do not assume A to be irreducible. The convergence question for iterative techniques continues to be studied extensively. The interest has been primarily in rearranging states on the rate of convergence. Here we begin the investigation of an alternate technique, namely a semi-iterative or averaging process, to attain convergence.

Key words. semi-iterative methods, iterative methods, singular M-matrices, $(C,1)$-summability, M-splittings

AMS(MOS) subject classifications. 15A06, 15A48, 60K20, 65F10

1. Introduction and notation. Many applications are modeled by a queueing system in which the problem is to find the stationary vector for a stochastic matrix. This is usually rewritten as

$$Ax = 0 \tag{1.1}$$

where A is a singular M-matrix. (See, e.g., Barker and Plemmons [1986], Kaufman [1983], Mitra and Tsoucas [1987], or Rose [1984].) An instance of such a model is the simple production line cited in Mitra and Tsoucas [1987], which is a tandem of M machines each of which is equipped with a finite buffer. Together with probabilistic assumptions concerning arrival rates and service times this is modeled by a finite-state continuous-time Markov process. By, for instance, observing the state of this process immediately after a state transition we proceed to a discrete-time Markov chain whose stationary vector we wish to compute.

Particular applications involve representing A as

$$A = I - B \tag{1.2}$$

where B is column stochastic, or as

$$A = D - L - U \tag{1.3}$$

where: D, L, and U are entrywise nonnegative; D is diagonal; and L and U are, respectively, strictly lower and strictly upper triangular. The solutions of (1.1) then correspond to solutions of

$$Bx = x \tag{1.2'}$$

or of

$$(D - L)^{-1}Ux = x. \tag{1.3'}$$

These are important special cases of a matrix splitting $A = M - N$ with a corresponding iteration matrix $T = M^{-1}N$. The iteration procedure becomes

$$Mx^{(k+1)} = Nx^{(k)}.$$

* Received by the editors March 23, 1987; accepted for publication July 29, 1987.

† Department of Mathematics, University of Missouri, Kansas City, Missouri 64110-2499.

‡ Permanent address, Department of Mathematics, Anhui University, Hefei, Anhui, People's Republic of China.

In general, if M is nonsingular and $T = M^{-1}N$ is semiconvergent, then for a suitable initial approximation $x^{(0)}$ the sequence

$$x^{(k)} = T^k x^{(0)}$$

will converge to a nonzero solution of (1.1). This follows immediately from

$$Ax^{(k)} = MT^k x^{(0)} - NT^k x^{(0)} = NT^{k-1}x^{(0)} - NT^k x^{(0)}.$$

Recent studies have assumed the irreducibility of A together with various other conditions which ensure that

$$\lim_{n \to \infty} T^n$$

exists. Here we impose conditions on A which guarantee the convergence of semi-iterative techniques and also the convergence of relaxation methods for not necessarily irreducible A.

We draw heavily on the theory of nonnegative matrices, and our basic references are Berman and Plemmons [1979, Chap. 2], Gantmacher [1959, Vol. II, Chap. 14], and Varga [1962, Chap. 2]. Specifically, a matrix $B = (b_{ij})$ is *nonnegative* ($B \geqq 0$) if and only if $b_{ij} \geqq 0$ for all i and j. Further, if $r(B)$ denotes the spectral radius of B, then we know that $B \geqq 0$ implies $r(B) \in \sigma(B)$, where $\sigma(B)$ denotes the spectrum of B. In this case this eigenvalue is called the Perron root of B and is denoted by $\rho(B)$. A matrix A (nonnegative or otherwise) is termed *reducible* if there is a permutation matrix P such that

$$P^T A P = \begin{bmatrix} A_1 & 0 \\ B & A_2 \end{bmatrix}$$

where A_1 and A_2 are nonempty square submatrices. If A is not reducible, it is called *irreducible*. The convergence rate of an iteration matrix is controlled by the parameter $\gamma(T)$ which is defined by

$$\gamma(T) = \max \{|\lambda|: \lambda \in \sigma(T), \lambda \neq 1\}.$$

A matrix T is called *semiconvergent* provided $\lim_{n \to \infty} T^n$ exists. Thus T is semiconvergent if and only if $r(T) \leqq 1$, $1 \in \sigma(T)$ implies its elementary divisors are linear, and $\gamma(T) < 1$.

A matrix of the form

$$A = sI - B$$

where $B \geqq 0$ and $s \geqq \rho(B)$ is called an *M-matrix*. If $s = \rho(B)$, then A is a *singular M-matrix*. An important special class of singular M-matrices are the *Q-matrices* (see Rose [1984] or Barker and Plemmons [1986]). A singular M-matrix $A = (a_{ij})$ is a Q-matrix if

$$\sum_{i=1}^{n} a_{ij} = 0, \qquad 1 \leqq j \leqq n. \tag{1.4}$$

Finally, let A be an $n \times n$ matrix with spectrum $\sigma(A)$, and let $\lambda \in \sigma(A)$. Recall that the index of λ is defined by the condition

$$\operatorname{ind}_\lambda (A) = \inf \{k: \ker (\lambda I - A)^k = \ker (\lambda I - A)^{k+1}\}$$

where $\ker A$ is the kernel (or null space) of the linear transformation A.

2. Semi-iterative methods. Following Varga [1962, Chap. 4] we use "semi-iterative" to describe an iterative technique together with an algebraic combination of these vector iterates.

DEFINITION 2.1. Let Z be an $n \times n$ matrix with spectral radius $\rho(Z)$. Z is said to satisfy *property* (E) if and only if for all $\lambda \in \sigma(Z)$, $|\lambda| = \rho(Z)$ implies that the elementary divisors of λ are all linear.

Remark 2.2. The notation arises since this property allows us to prove a type of the mean ergodic theorem. This property is obtained for important classes of matrices. It is classical that this property is satisfied if A is a stochastic matrix. Further, if $A = M - N$ is a nontrivial M-splitting of an irreducible M-matrix A (that is, M is itself an M-matrix) then from Theorem 3.5 of Schneider [1984] we know that the iteration matrix $T = M^{-1}N$ satisfies property (E). Quite similar ideas for semiconvergent iteration matrices T are discussed in Neumann and Plemmons [1978].

LEMMA 2.3. *Let $Z \geqq 0$ be an $n \times n$ matrix which satisfies property (E) and for which $\rho(Z) = 1$. Then*

$$\frac{1}{n+1}\sum_{k=0}^{n} Z^k$$

converges to a projection $P \geqq 0$. Further, if $x^{(0)}$ is any vector, then

$$Z(Px^{(0)}) = Px^{(0)}.$$

Proof. The proof is standard (cf. Barker [1974]), so it will only be outlined here. Choose S so that $S^{-1}ZS$ is in Jordan normal form:

$$S^{-1}ZS = I \oplus \omega_1 I \oplus \cdots \oplus \omega_p I \oplus (\lambda_1 I + U) \oplus \cdots \oplus (\lambda_q I + U)$$

where the I's denote identity matrices of appropriate sizes, the U's have 1's along the first superdiagonal and zeros elsewhere, $|\omega_j| = 1, j = 1, \cdots, p$ and $|\lambda_j| < 1, \lambda = 1, \cdots, q$. It is well known that

$$\frac{1}{n+1}\left|\sum_{k=0}^{n} \omega^k\right| \to 0$$

as $n \to \infty$ when $|\omega| \leqq 1$, $\omega \neq 1$. Further,

$$\sum_{k=0}^{n} (\lambda_s I + U)^k$$

is bounded. Thus,

$$\lim_{n\to 0}\left(\frac{1}{n+1}\sum_{k=0}^{n}(S^{-1}ZS)^k\right) = I \oplus 0 \oplus \cdots \oplus 0.$$

Consequently, the limit

$$\lim_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^{n} Z^k = P$$

is a nonnegative projection which commutes with Z. Finally, for any vector x we have

$$Z\left(\frac{1}{n+1}\sum_{k=0}^{n} Z^k\right)x = \frac{1}{n+1}(Z+\cdots+Z^{n+1})x$$

$$= \frac{1}{n+2}\left(\frac{n+2}{n+1}\right)(I+Z+\cdots+Z^{n+1})x - \frac{1}{n+2}\left(\frac{n+2}{n+1}\right)x,$$

and a passage to the limit as $n \to \infty$ shows that $ZPx = Px$.

Remark. If x has all entries positive, then since $P \geqq 0$ it follows that $Px \neq 0$. Thus for a positive vector x we see that Px is a nonnegative eigenvector of Z belonging to 1.

Semi-iterative methods.

(a) If $A = I - B$ we call the scheme

$$x^{(n)} = Bx^{(n-1)},$$

$$y^{(n)} = \frac{1}{n+1}(x^{(n)}+\cdots+x^{(0)}) = \frac{1}{n+1}x^{(n)} + \frac{n}{n+1}y^{(n-1)}$$

the $(C,1)$-*Jacobi method.*

(b) If $A = D - L - U$ we call either of the schemes

$$(D-L)x^{(n)} = Ux^{(n-1)},$$

$$y^{(n)} = \frac{1}{n+1}x^{(n)} + \frac{n}{n+1}y^{(n-1)},$$

or

$$(D-U)x^{(n)} = Lx^{(n-1)},$$

$$y^{(n)} = \frac{1}{n+1}x^{(n)} + \frac{n}{n+1}y^{(n-1)}$$

a $(C,1)$-*Gauss–Seidel method.*

THEOREM 2.4. *Let A be an $n \times n$ singular M-matrix.*

(a) *If $A = I - B$ where B is a stochastic matrix, then for any positive initial vector $x^{(0)}$ the $(C,1)$-Jacobi method converges to a nonnegative nonzero solution of $Ax = 0$.*

(b) *If $A = D - L - U$ is an irreducible singular M-matrix as in* (1.3), *then for any positive initial vector $x^{(0)}$ the $(C,1)$-Gauss–Seidel method converges to a nonnegative nonzero solution of $Ax = 0$ which is unique up to scalar multiples.*

Remark. Although the limit in (b) is unique apart from scalar multiples, the same is not generally true in (a). In fact, if we decompose $\mathbb{R}^n$ as

$$\mathbb{R}^n = E \oplus S,$$

where E is the eigenspace corresponding to $1 \in \sigma(B)$ and S is the invariant subspace corresponding eigenvalues of B, then the $(C,1)$-limit

$$\lim_{n\to\infty} \frac{1}{n+1}\sum_{k=0}^{n} B^k = P$$

is the projection onto E along S. P is nonnegative and is the operator residue of the resolvent $(zI - B)^{-1}$ at $z = 1$. Thus, the $(C,1)$-limit which is $Px^{(0)}$ depends upon $x^{(0)}$ and is not unique (up to scalar multiples) unless $\dim E = 1$. Since B is stochastic, $\dim E = 1$ when and only when B is irreducible.

Proof. (a) It is well known (cf. Gantmacher [1959, II, p. 86]) that if B is stochastic, then B satisfies property (E). Now apply Lemma 2.3.

(b) If A is irreducible, then $A = (D - L) - U$ is a nontrivial M-splitting. Put $T = (D - L)^{-1}U$. From Theorem 3.5 of Schneider [1984] we have that T satisfies property (E) and the result follows from the lemma.

It is well known (see Gantmacher [1959]) that if B is an irreducible nonnegative matrix, then

$$B = \rho(B)D^{-1}PD$$

where D is a positive diagonal matrix and P is stochastic. Thus, in principle, case (b) can be reduced to case (a) by a diagonal similarity and a scaling. The second case is stated for comparison with previous work. We know (cf. Barker and Plemmons [1986, p. 395]) that for an irreducible singular M-matrix we can always permute the rows and columns so that the Gauss–Seidel method converges. The comparison to be made is between the cost of the graph search together with the Gauss–Seidel iterations and the cost of averaging for a $(C,1)$-Gauss–Seidel method.

The ideas of $(C,1)$ iterative solutions are by no means new. They are discussed extensively in Rothblum ([1980] and [1981]), who uses the term average convergent. Rothblum obtains explicit forms of solutions in terms of Drazin inverses and relates $(C,1)$-convergence of nonnegative matrices to the structure of the fundamental classes. In particular Lemma 2.3 is contained in his Lemma 3.2. Our primary observation here is that $(C,1)$-convergence applies to some important cases where we can spell out a (perhaps) useful iterative method.

3. Basic methods and singular *M*-matrices. A great deal is known about irreducible M-matrices. In particular, one may consult Schneider [1984] and Berman and Plemmons [1979] from which we quote the following result.

THEOREM 3.1. *Let A be a singular irreducible M-matrix of order n. Then*

(a) rank $A = n - 1$,

(b) *there is a positive vector x such that $Ax = 0$,*

(c) *A has property C, that is, $A = sI - B$, $s > 0$, $B \geqq 0$, and $(1/s)B$ is semiconvergent.*

If A is an irreducible singular M-matrix, then there are positive diagonal matrices D_1 and D_2 such that

$$D_1AD_2 = I - L - U = I - B$$

where B is column stochastic and trace $B = 0$. In fact, for each $\varepsilon > 0$ there is a positive diagonal matrix

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

such that all the column sums of $D(\varepsilon)(A + \varepsilon I)$ are positive:

$$(a_{jj} + \varepsilon)\, d_j(\varepsilon) > \sum_{i \neq j} |a_{ij}|\, d_i(\varepsilon), \qquad j = 1, \cdots, n. \tag{3.2}$$

We may divide both sides of (3.2) by $(d_1(\varepsilon)^2 + \cdots + d_n(\varepsilon)^2)^{1/2}$ and so may take the vector $(d_1(\varepsilon), \cdots, d_n(\varepsilon))$ to be on the unit sphere. By compactness there is a subsequence $\varepsilon_k \to 0$ such that $d_j(\varepsilon_k) \to d_j \geqq 0$ for $j = 1, \cdots, n$. Hence

$$a_{jj}d_j \geqq \sum_{i \neq j} |a_{ij}|\, d_i, \qquad j = 1, \cdots, n. \tag{3.3}$$

Since $(d_1, \cdots, d_n)$ is a unit vector it is nonzero and so

$$S = \{j : d_j > 0\} \neq \varnothing.$$

We may assume without loss that

$$S=\{1,2,\cdots,l\}, \qquad l\leqq n.$$

If $l<n$, then for each j with $l<j\leqq n$ and all $i<j$,

$$a_{ij}=0.$$

This follows from (3.3). Since this contradicts the irreducibility of A, we have that $l=n$. Finally, let

$$d=\operatorname{diag}(d_1,\cdots,d_n).$$

Then for $D_1=D^{-1}$ and $D_2=D\operatorname{diag}(1/a_{11},\cdots,1/a_{nn})$, we have

$$D_1AD_2=I-B$$

where $B\geqq 0$, trace $B=0$, and B is column stochastic.

The hypothesis of irreducibility also imposes restrictions on the multiplicity and index of the eigenvalue 0 of A. We shall deal with a slightly larger class of singular M-matrices which contains both the scaled irreducible M-matrices and the Q-matrices. Specifically, we assume that

$$A=I-L-U=I-B \tag{3.4}$$

is a singular M-matrix, L is a strictly lower triangular matrix, while U is (not necessarily strictly) upper triangular and

$$\operatorname{ind}_0(A)=1. \tag{3.5}$$

Note that A is a Q-matrix when B is column stochastic.

If $A=I-B$ is irreducible, then B is irreducible. Thus for any $\alpha>0$, $\alpha I+B$ is primitive. Split A as

$$A=\frac{1}{\varepsilon}I-\left(\frac{1-\varepsilon}{\varepsilon}I+B\right)$$

and denote the JOR iteration matrix by

$$W(\varepsilon)=\left(\frac{1}{\varepsilon}I\right)^{-1}\left(\frac{1-\varepsilon}{\varepsilon}I+B\right)=(1-\varepsilon)I+\varepsilon B \tag{3.6}$$

for $0<\varepsilon<1$. Young [1972] calls ε the Jacobi overrelaxation (JOR) parameter while others (e.g., Mitra and Tsoucas [1987]) use $\alpha=1-\varepsilon$ ($0\leqq\alpha<1$) as the relaxation parameter. $W(\varepsilon)$ is primitive and $\rho(W(\varepsilon))=1$, whence $W(\varepsilon)$ is semiconvergent. In particular, if $x^{(0)}\geqq 0$ is a nonzero initial vector, then the sequence

$$x^{(k)}=[W(\varepsilon)]^kx^{(0)}$$

converges to a positive solution of $Ax=0$.

There is an analogous Gauss–Seidel type of iteration. For let $0<\varepsilon<1$ and put

$$A=I-L-U=\left(\frac{1}{\varepsilon}I-L\right)-\left(\frac{1-\varepsilon}{\varepsilon}I+U\right).$$

We then obtain the successive overrelaxation (SOR) iteration matrix (cf. Barker and Plemmons [1986, p. 395])

$$T(\varepsilon)=\left(\frac{1}{\varepsilon}I-L\right)^{-1}\left(\frac{1-\varepsilon}{\varepsilon}I+U\right)=(I-\varepsilon L)^{-1}((1-\varepsilon)I+\varepsilon U). \tag{3.7}$$

We now have

$$\begin{aligned} T(\varepsilon) &= (I-\varepsilon L)^{-1}((1-\varepsilon)I+\varepsilon U) \\ &= (I+\varepsilon L+\cdots+\varepsilon^{n-1}L^{n-1})((1-\varepsilon)I+\varepsilon U) \\ &\geqq \varepsilon(1-\varepsilon)(I+L+U) \\ &= \varepsilon(1-\varepsilon)(I+B). \end{aligned}$$

If A is irreducible, then $I + B$ and hence $T(\varepsilon)$ are primitive. Since $\rho(T(\varepsilon)) = 1$, we see that $T(\varepsilon)$ is semiconvergent. If $x^{(0)} \geqq 0$ ($x^{(0)} \neq 0$), then

$$x^* = \lim [T(\varepsilon)]^k x^{(0)}$$

is a positive solution of (1.1).

We now relax the conditions on A in the convergence results.

PROPOSITION 3.8. *Let A be an $n \times n$ matrix which satisfies the assumptions* (3.4) *and* (3.5). *Then the iteration scheme*

$$x^{(k+1)} = W(\varepsilon)x^{(k)}, \qquad x^{(0)} > 0,$$

converges for any ε in (0, 1) *to a nonzero nonnegative solution of* (1.1).

Proof. Clearly $\sigma(W(\varepsilon)) = \{1-\varepsilon+\varepsilon\lambda: \lambda \in \sigma(B)\}$. We have $\rho(W(\varepsilon)) = 1$, this eigenvalue has index 1, and $\gamma(W(\varepsilon)) < 1$. Therefore $W(\varepsilon)$ is semiconvergent for $0 < \varepsilon < 1$.

Conditions (3.4) and (3.5) are slightly weaker than the assumption that the iteration matrix is semiconvergent. The latter assumption has been used for an extensive study of regular splittings in, for instance, Neumann and Plemmons [1978]. The main result of our paper is the next theorem which shows that convergence of the iteration scheme occurs with our somewhat weaker hypotheses.

THEOREM 3.9. *Let A be an $n \times n$ matrix which satisfies the assumptions* (3.4) *and* (3.5). *Then the iteration scheme*

$$x^{(k+1)} = T(\varepsilon)x^{(k)}, \qquad x^{(0)} > 0,$$

converges to a nonnegative nonzero solution of (1.1).

Proof. As in the discussion following (3.7), we have

$$T(\varepsilon) = (I+\varepsilon L+\cdots+\varepsilon^{n-1}L^{n-1})((1-\varepsilon)I+\varepsilon U). \tag{3.10}$$

From (3.10) we infer that

$$\begin{aligned} [T(\varepsilon)]^k &\leqq [(I+\varepsilon+\cdots+\varepsilon^{n-1}L^{n-1})(I+\varepsilon U)]^k \\ &\leqq [I+\varepsilon L+\varepsilon U]^{nk} \leqq \left[\frac{W(\varepsilon)}{1-\varepsilon}\right]^{nk}, \end{aligned} \tag{3.11}$$

and that

$$T(\varepsilon) \geqq (1-\varepsilon)I+\varepsilon(1-\varepsilon)L+\varepsilon U \geqq (1-\varepsilon)\,W(\varepsilon). \tag{3.12}$$

For a suitable permutation matrix P, we have

$$PT(\varepsilon)P^{-1} = \begin{bmatrix} T_{11} & 0 & \cdots & 0 \\ T_{21} & T_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ T_{m1} & T_{m2} & \cdots & T_{mm} \end{bmatrix},$$

where each T_{ll} ($l = 1, \cdots, m$) is an irreducible square matrix. (N.B. The 1×1 zero matrix is regarded as an irreducible matrix.) If $T(\varepsilon)$ is not semiconvergent since by Theorem 4.5 of Schneider [1984] we have

$$\operatorname{ind}_1 (T(\varepsilon)) = \operatorname{ind}_0 (A) = 1,$$

then there is some T_{ll} and some permutation matrix Q for which we have

$$\rho(T_{ll}) = \rho(T(\varepsilon)) = 1$$

and

$$QT_{ll}Q^{-1} = \begin{bmatrix} 0 & C_{12} & 0 & \cdots & 0 \\ 0 & 0 & C_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & C_{hh-1} \\ C_{h1} & 0 & 0 & \cdots & 0 \end{bmatrix} = C \tag{3.13}$$

where $h > 1$. Now let

$$C^k = [c_{ij}^{(k)}], \qquad 1 \leqq i, \quad j \leqq g, \quad k = 1, 2, \cdots$$

where g is of the order T_{ll}. It is clear from (3.13) that for any pair (i, j) there is a k, $1 \leqq k \leqq h$, such that

$$c_{ij}^{(k)} = 0,$$

and, further,

$$c_{ij}^{(k+fh)} = 0, \qquad f = 0, 1, 2, \cdots.$$

We have by (3.12) that

$$P[T(\varepsilon)]^k P^T \geqq (1-\varepsilon)^k P[W(\varepsilon)]^k P$$

and hence

$$(1-\varepsilon)^{-2k} C^k \geqq Q \frac{E^{(k)}}{(1-\varepsilon)^k} Q^T \tag{3.14}$$

where $E^{(k)}$ is the $g \times g$ submatrix of $P[W(\varepsilon)]^k P^T$ corresponding to T_{ll}. (Note that the exponent in $E^{(k)}$ is an index, not a power.) Now $W(\varepsilon)$ is semiconvergent so that

$$\lim_{k \to \infty} P[W(\varepsilon)]^k P^T$$

exists and, consequently,

$$\lim_{k \to \infty} E^{(k)}$$

exists. But by choosing a suitable subsequence we see that $c_{ij}^{(k)} = 0$, whence

$$\lim [E^{(k)}(1-\varepsilon)^{-k}] = 0.$$

But then from (3.11) it follows that

$$\lim_{k \to \infty} (T_{ll})^k = 0,$$

which contradicts $\rho(T_{ll}) = 1$. Therefore, $T(\varepsilon)$ is semiconvergent.

4. Rates of convergence. Given a convergent iteration scheme

$$x^{(k+1)} = Nx^{(k)}, \qquad k = 0, 1, \cdots \tag{4.1}$$

the asymptotic rate of convergence is the parameter $-\ln[\gamma(N)]$ (see, e.g., Funderlic and Plemmons [1984]). Thus, the smaller $\gamma(N)$ the faster we expect (4.1) to converge.

Let $N(\varepsilon)$ in (4.1) denote either the JOR or the SOR iteration matrices, that is,

$$N(\varepsilon) = W(\varepsilon) = (1-\varepsilon)I + \varepsilon B \qquad (0 < \varepsilon \leqq 1)$$

or

$$N(\varepsilon) = T(\varepsilon) = \left(\frac{1}{\varepsilon}I - L\right)^{-1}\left(\frac{1-\varepsilon}{\varepsilon}I + U\right) \qquad (0 < \varepsilon \leqq 1).$$

The entries of $N(\varepsilon)$ are continuous functions of ε so that $\gamma(N(\varepsilon))$ is a continuous function of ε.

DEFINITION 4.2. We call $\varepsilon_0 \in (0, 1]$ an *optimal value* if

$$\gamma(N(\varepsilon_0)) \leqq \gamma(N(\varepsilon))$$

for each $\varepsilon \in (0, 1]$.

Considerable work has been done on these optimal values. For a general survey including the complex case, see Hadjidimos [1984], [1985].

If there is an optimal value ε_0, then $N(\varepsilon_0)$ should give the fastest rate of convergence in the iteration (4.1). The next result shows that the optimal values exist.

THEOREM 4.3. *For the basic iteration schemes an optimal value ε_0, as defined above, always exists.*

Proof. Clearly $W(0) = I$. If we rewrite $T(\varepsilon)$ as

$$T(\varepsilon) = \varepsilon(I - \varepsilon L)^{-1}\left(\frac{1-\varepsilon}{\varepsilon}I + U\right) = (I - \varepsilon L)^{-1}((1-\varepsilon)I + \varepsilon U),$$

then $T(0) = I$. Since

$$\lim_{\varepsilon \to 0^+} N(\varepsilon) = I$$

and 1 is an eigenvalue of multiplicity n of I, we have

$$\lim_{\varepsilon \to 0^+} \gamma(N(\varepsilon)) = 1.$$

For simplicity let $g(\varepsilon) = \gamma(N(\varepsilon))$. Then $g(\varepsilon)$ is continuous on $[0, 1]$ so there exists a $\delta_1 > 0$ such that

$$g(\varepsilon) \geqq g(\delta_1) \tag{4.4}$$

for any $0 < \varepsilon < \delta_1 < 1$. If $N(1)$ is semiconvergent, there is an $\varepsilon_0 \in [\delta_1, 1] \subset (0, 1]$ at which g takes a minimum value. Then (4.4) implies that $g(\varepsilon_0)$ is a minimum on $(0, 1]$, that is, ε_0 is an optimal value.

If $N(1)$ is not semiconvergent, then since $\text{ind}_1(N(1)) = 1$ it follows that $N(1)$ has at least two different eigenvalues on the unit circle. In this case

$$\lim_{\varepsilon \to 1^-} \gamma(N(\varepsilon)) = 1,$$

whence there is a $\delta_2 > 0$ such that

$$g(\varepsilon) \geqq g(1 - \delta_2) \tag{4.5}$$

for any ε in $(1-\delta_2, 1)$. Since there is an $\varepsilon_0 \in [\delta_1, 1-\delta]$ at which g takes a minimum value, (4.4) and (4.5) imply that this ε_0 is an optimal value.

The computation of the optimal value ε_0 seems to be a difficult problem. As an indication of how we might approach the general case, we shall consider the 3×3 case and obtain an estimate for ε_0. Here A is a singular M-matrix of rank 2 whose column sums are zero. That is, $A = I - B$ is a Q matrix of rank 2 and trace 3.

In detail, let

$$A = \begin{bmatrix} 1 & t_2-1 & -t_3 \\ -t_1 & 1 & t_3-1 \\ t_1-1 & -t_2 & 1 \end{bmatrix} = I-L-U \tag{4.6}$$

where

$$L = \begin{bmatrix} 0 & 0 & 0 \\ t_1 & 0 & 0 \\ 1-t_1 & t_2 & 0 \end{bmatrix}, \qquad U = \begin{bmatrix} 0 & 1-t_2 & t_3 \\ 0 & 0 & 1-t_3 \\ 0 & 0 & 0 \end{bmatrix}.$$

We have $0 \leqq t_i \leqq 1$, $i = 1, 2, 3$, and $\det A = 0$. This last condition yields

$$1-(1-t_1)(1-t_2)(1-t_3)-t_1t_2t_3 = t_1(1-t_2)+t_2(1-t_3)+t_3(1-t_1).$$

We have

$$\begin{aligned} T(\varepsilon) &= \left(\frac{1}{\varepsilon}I-L\right)^{-1}\left(\frac{1-\varepsilon}{\varepsilon}I+U\right) \\ &= \begin{bmatrix} \varepsilon & 0 & 0 \\ t_1\varepsilon & \varepsilon & 0 \\ t_1t_2\varepsilon^3+(1-t_1)\varepsilon^2 & t_2\varepsilon & \varepsilon \end{bmatrix} \begin{bmatrix} \dfrac{1-\varepsilon}{\varepsilon} & 1-t_2 & t_3 \\ 0 & \dfrac{1-\varepsilon}{\varepsilon} & 1-t_3 \\ 0 & 0 & \dfrac{1-\varepsilon}{\varepsilon} \end{bmatrix}. \end{aligned} \tag{4.7}$$

Let $\sigma(T(\varepsilon)) = \{1, \lambda_+(\varepsilon), \lambda_-(\varepsilon)\}$; then

$$\begin{aligned} \lambda_\pm(\varepsilon) &= \tfrac{1}{2}\{\operatorname{tr} T(\varepsilon) - 1 \pm [(\operatorname{tr} T(\varepsilon)-1)^2 - 4 \det T(\varepsilon)]^{1/2}\} \\ &= \tfrac{1}{2}(\phi(\varepsilon) \pm \sqrt{\psi(\varepsilon)}) \end{aligned} \tag{4.8}$$

where for $b = t_1t_2t_3$ we have

$$\phi(\varepsilon) = \operatorname{tr} T(\varepsilon) - 1 = b\varepsilon^3 + [1-b-(1-t_1)(1-t_2)(1-t_3)]\varepsilon^2 - 3\varepsilon + 2$$

and

$$\psi(\varepsilon) = [\phi(\varepsilon)]^2 - 4(1-\varepsilon)^3.$$

LEMMA 4.9. *If* $\psi(\varepsilon_\alpha) = \psi(\varepsilon_\beta) = 0$, *for* $0 \leqq \varepsilon_\alpha < \varepsilon_\beta < 1$ *and* $|\psi(\varepsilon)| > 0$ *on* $(\varepsilon_\alpha, \varepsilon_\beta)$, *then*

$$\min_{\varepsilon_\alpha \leqq \varepsilon \leqq \varepsilon_\beta} g(\varepsilon) \leqq g(\varepsilon^*) \tag{4.10}$$

where ε^* *is the largest zero of* $\psi(\varepsilon)$ *on* (0, 1).

Proof. There are two cases.

Case A. $\psi(\varepsilon) < 0$ on $(\varepsilon_\alpha, \varepsilon_\beta)$. In this case (4.10) follows from

$$g(\varepsilon) = \gamma(T(\varepsilon)) = \tfrac{1}{2}[(\phi(\varepsilon))^2 + 4(1-\varepsilon)^3 - (\phi(\varepsilon))^2]^{1/2} = (1-\varepsilon)^{3/2}$$

and

$$g'(\varepsilon) = -\tfrac{3}{2}(1-\varepsilon)^{1/2} < 0 \quad \text{on } (0, 1).$$

Case B. $\psi(\varepsilon) > 0$ on $(\varepsilon_\alpha, \varepsilon_\beta)$. In this case we have

$$g(\varepsilon) = \tfrac{1}{2}[|\phi(\varepsilon)| + \sqrt{\psi(\varepsilon)}], \qquad \varepsilon \in (\varepsilon_\alpha, \varepsilon_\beta)$$

and

$$g'(\varepsilon) = [\phi'(\varepsilon)(\operatorname{sgn}(\phi(\varepsilon))\sqrt{\psi(\varepsilon)} + \phi(\varepsilon)) + 6(1-\varepsilon)^2]/(2\sqrt{\psi(\varepsilon)}).$$

We claim that

$$\phi'(\varepsilon) = 3b\varepsilon^2 + 2[t_1(1-t_2) + t_2(1-t_3) + t_3(1-t_1)]\varepsilon - 3 < 0 \tag{4.11}$$

for $\varepsilon \in (0, 1)$. Obviously this holds when $b = 0$. When $b > 0$, $\phi''(\varepsilon)$ has only the zero

$$\varepsilon = -\frac{1}{3b}(t_1(1-t_2) + t_2(1-t_3) + t_3(1-t_1)) < 0,$$

whence

$$\max_{0 \leqq \varepsilon \leqq 1} \phi'(\varepsilon) = \max\{\phi'(0), \phi'(1)\} < 0,$$

and this implies (4.11).

Since $\phi(\varepsilon) = [\phi(\varepsilon)]^2 - 4(1-\varepsilon)^3 > 0$ implies $|\phi(\varepsilon)| \geqq \sqrt{\psi(\varepsilon)} > 0$ on $(\varepsilon_\alpha, \varepsilon_\beta)$, then $\phi(\varepsilon)$ is of constant sign on this interval and its sign is also the sign of $\operatorname{sgn}(\phi(\varepsilon))\sqrt{\psi(\varepsilon)} + \phi(\varepsilon)$.

If $\phi(\varepsilon) < 0$ on $(\varepsilon_\alpha, \varepsilon_\beta)$, then $g'(\varepsilon) > 0$ there and so

$$\min_{\varepsilon_\alpha \leqq \varepsilon \leqq \varepsilon_\beta} g(\varepsilon) = g(\varepsilon_\alpha) \leqq g(\varepsilon^*).$$

If $\phi(\varepsilon) > 0$ on $(\varepsilon_\alpha, \varepsilon_\beta)$, then

$$g(\varepsilon) = \tfrac{1}{2}[\phi(\varepsilon) + \sqrt{\psi(\varepsilon)}] \geqq \tfrac{1}{2}\phi(\varepsilon_\beta) = g(\varepsilon_\beta).$$

Therefore

$$\min_{\varepsilon_\alpha \leqq \varepsilon \leqq \varepsilon_\beta} g(\varepsilon) = g(\varepsilon_\beta) \leqq g(\varepsilon^*).$$

THEOREM 4.12. *Let* $N = T(\varepsilon)$ *be the* 3×3 *matrix* (4.7); *then when*

$$(1-t_1)(1-t_2)(1-t_3) = 0$$

the iteration scheme (4.1) *has the only optimal value one*; *when*

$$(1-t_1)(1-t_2)(1-t_3) \neq 0$$

it has only the optimal value ε^*, *the largest zero of* $\psi(\varepsilon)$ *on* (0, 1).

Proof. $\psi(0) = 0$. If $(1-t_1)(1-t_2)(1-t_3) = 0$, then we have

$$\phi(\varepsilon) = (\varepsilon - 1)(b\varepsilon^2 + \varepsilon - 2),$$

$$\psi(\varepsilon) = \varepsilon^2(\varepsilon-1)^2(b^2\varepsilon^2 + 2b\varepsilon + 1 - 4b)$$

and hence $\psi(1) = 0$. When $b \leqq \frac{1}{4}$, we have $\phi(\varepsilon) > 0$ and $\psi(\varepsilon) > 0$ on (0, 1). In this case $g(\varepsilon) \geqq \frac{1}{2}\phi(1) = g(1)$ by (4.13), that is,

$$\min_{0 \leqq \varepsilon \leqq 1} g(\varepsilon) = g(1).$$

Lemma 4.9 implies that $g(\varepsilon^*) = \min_{0 \leq \varepsilon \leq \varepsilon^*} g(\varepsilon)$ whenever ε^* exists. When $b > \frac{1}{4}$,

$$\varepsilon^* = \frac{1}{b}(2\sqrt{b} - 1) \text{ exists and } \phi(\varepsilon) > 0,\ \psi(\varepsilon) > 0 \text{ on } (\varepsilon^*, 1).$$

In this case

$$g(1) = \min_{\varepsilon^* \leq \varepsilon \leq 1} g(\varepsilon) = \min_{0 \leq \varepsilon \leq 1} g(\varepsilon).$$

If $(1 - t_1)(1 - t_2)(1 - t_3) \neq 0$, then $\phi(1) = -(1 - t_1)(1 - t_2)(1 - t_3) < 0$ and $\phi(0) = 2 > 0$. So there exists an $\varepsilon' \in (0, 1)$ such that $\phi(\varepsilon') = 0$ and $\psi(\varepsilon') < 0$. Since $\psi(1) = |\phi(1)| > 0$, there must be an $\varepsilon_1 \in (0, 1)$ such that $\phi(\varepsilon_1) = 0$. In this case ε^* exists and $\phi(\varepsilon) < 0$, $\psi(\varepsilon) > 0$ on $(\varepsilon^*, 1)$. Therefore $g'(\varepsilon) > 0$ on $(\varepsilon^*, 1)$, $\min_{\varepsilon^* \leq \varepsilon \leq 1} g(\varepsilon) = g(\varepsilon^*)$, and hence

$$g(\varepsilon^*) = \min_{0 \leq \varepsilon \leq 1} g(\varepsilon).$$

Note that if $(1 - t_1)(1 - t_2)(1 - t_3) = 0$, then $T(1)$ is always semiconvergent.

As an example for the matrix

$$A = \begin{bmatrix} 1 & -\frac{1}{3} & -\frac{2}{3} \\ -\frac{2}{3} & 1 & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{2}{3} & 1 \end{bmatrix}$$

the optimal value is $\varepsilon^* = 0.965$ with $g(\varepsilon^*) = 0.00396$, $g(1) = 0.037$, and $g(0^+) = 1$.

5. Summary. We have discussed some of the theory concerning the iterative solution of

$$Ax = 0$$

when A is a singular M-matrix. Most applications to date can be reduced to the situation where A is irreducible. However, we show that under certain conditions (cf. Theorem 2.4) that a $(C, 1)$-semi-iterative procedure converges to a nonzero solution even for reducible A. The comparison for an irreducible $A = I - L - U$ is between the cost of the averages in the $(C, 1)$-process and the cost of a graph algorithm to determine a permutation P such that P^TAP has a semiconvergent iteration matrix. The main result is that if

$$A = I - L - U$$

is a singular M-matrix and if the index of 0 in A is 1, then the SOR iteration converges. Irreducibility of A implies our condition, but the two are not equivalent. The computation of the optimal value of the relaxation parameter seems to be difficult. We treat an instance of the 3×3 case and obtain some estimates.

REFERENCES

G. P. BARKER [1974], *Stochastic matrices over cones*, Linear and Multilinear Algebra, 1, pp. 279–287.

G. P. BARKER AND R. J. PLEMMONS [1986], *Convergent iterations for computing stationary distributions of Markov chains*, SIAM J. Algebraic Discrete Methods, 7, pp. 390–398.

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

R. E. FUNDERLIC AND R. J. PLEMMONS [1984], *A combined direct-iterative method for certain M-matrix linear systems*, SIAM J. Algebraic Discrete Methods, 5, pp. 33–42.

F. R. GANTMACHER [1959], *Matrix Theory*, Vol. II, Chelsea, New York.

A. HADJIDIMOS [1984], *The optimal solution to the problem of complex extrapolation of a first-order scheme*, Linear Algebra Appl., 62, pp. 241–261.

——— [1985], *On the optimization of the classical iterative schemes for the solution of complex singular linear systems*, SIAM J. Algebraic Discrete Methods, 6, pp. 555–566.

L. KAUFMAN [1983], *Matrix methods for queueing problems*, SIAM J. Sci. Statist. Comput., 4, pp. 525–552.

D. MITRA AND P. TSOUCAS [1987], *Relaxations for the numerical solutions of some stochastic problems*, to appear.

M. NEUMANN AND R. J. PLEMMONS [1978], *Convergent nonnegative matrices and iterative methods for consistent linear systems*, Numer. Math., 31, pp. 173–186.

D. J. ROSE [1984], *Convergent regular splittings for singular M-matrices*, SIAM J. Algebraic Discrete Methods, 5, pp. 133–144.

U. G. ROTHBLUM [1980], *Convergence properties of powers of matrices with applications to iterative methods for solving linear systems*, in External Methods and Systems Analysis, A. V. Fiacco and K. O. Kortanek, eds. Lecture Notes in Economics and Mathematical Systems 174, Springer-Verlag, Berlin, New York, pp. 231–247.

——— [1981], *Resolvents expansions of matrices and applications*, Linear Algebra Appl., 38, pp. 33–49.

H. SCHNEIDER [1984], *Theorems on M-splittings of a singular M-matrix which depend on graph structure*, Linear Algebra Appl., 58, pp. 407–424.

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

D. M. YOUNG [1972], *Iterative Solutions of Large Linear Systems*, Academic Press, New York.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

005

TOEPLITZ SYSTEMS ASSOCIATED WITH THE PRODUCT OF A FORMAL LAURENT SERIES AND A LAURENT POLYNOMIAL*

WILLIAM F. TRENCH†

Abstract. A method is proposed for solving linear algebraic systems with Toeplitz matrices generated by $T(z) = C(z)\Phi(z)$, where $C(z)$ is a Laurent polynomial and $\Phi(z)$ is a formal Laurent series, and a convenient method is available for solving systems with Toeplitz matrices generated by $\Phi(z)$. Special cases of the method provide $O(n)$ procedures for solving $n \times n$ systems with banded or rationally generated Toeplitz matrices. The latter do not require recursion with respect to n.

Key words. Toeplitz systems, banded Toeplitz matrices, rationally generated Toeplitz matrices

AMS(MOS) subject classifications. 15A06, 65F05

1. Introduction. To motivate the problem considered here, let $\{x_j\}$ be a wide-sense stationary time series (possibly complex-valued) with zero mean and covariance $E(x_i\bar{x}_j) = \phi_{i-j}$. If

$$y_j = \sum_{l=0}^{p} b_l x_{j-l}, \qquad -\infty < j < \infty,$$

then $\{y_j\}$ has zero mean and covariance $E(y_i\bar{y}_j) = t_{i-j}$, where

$$t_i = \sum_{l=-p}^{p} c_l \phi_{i-l}, \tag{1}$$

with

$$c_l = \sum_{\nu=0}^{p-l} \bar{b}_\nu b_{\nu+l}, \qquad 0 \leqq l \leqq p,$$

and

$$c_l = \sum_{\nu=0}^{p+l} b_\nu \bar{b}_{\nu-l}, \qquad -p \leqq l \leqq -1.$$

Minimum variance estimation problems concerning the time series $\{y_j\}$ require solutions of the systems

$$T_n X = Y, \tag{2}$$

where T_n is the $n \times n$ Toeplitz matrix

$$T_n = (t_{i-j})_{i,j=1}^{n}. \tag{3}$$

(See, e.g., [16, pp. 20–23].) Definition (1) suggests that if we have an efficient way to solve the systems

$$\Phi_m U = V, \tag{4}$$

where

$$\Phi_m = (\phi_{i-j})_{i,j=1}^{m}, \tag{5}$$

* Received by the editors November 15, 1985; accepted for publication (in revised form) August 3, 1987.
† Department of Mathematics, Trinity University, San Antonio, Texas 78284.

then it should be possible to exploit it in solving (2). Here we propose a method that does this; however, since our results are not restricted to systems with positive definite Hermitian Toeplitz matrices, we first formulate the situation more generally.

Let

$$\Phi(z)=\sum_{j=-\infty}^{\infty}\phi_j z^j$$

be a formal Laurent series, and let

$$C(z)=\sum_{j=-q}^{p}c_j z^j \tag{6}$$

be a Laurent polynomial, with

$$p,q\geqq 0,\quad p+q=k\geqq 1,\quad c_p c_{-q}\neq 0. \tag{7}$$

Now define

$$T(z)=C(z)\Phi(z)=\sum_{j=-\infty}^{\infty}t_j z^j,$$

so that

$$t_i=\sum_{l=-q}^{p}c_l\phi_{i-l}. \tag{8}$$

We are still interested in solving (2).

There are many algorithms for solving Toeplitz systems that take advantage of their special simplicity. (See, e.g., [3], [8], [11], [12], [17] and [18]—by no means a complete list.) However, most require assumptions that are not met by all Toeplitz matrices, and some are stable only for certain classes of Toeplitz matrices. (In this connection, see [2].) Our results should be useful if there is a convenient algorithm for dealing with the matrices generated by $\Phi(z)$ which does not apply to those generated by $T(z)$. This could be so, for example, if the former are Hermitian, symmetric, triangular, or positive definite, or if there is a convenient explicit formula for their inverses, while the latter do not exhibit the desirable property. Our results provide a way to transfer the burden of computation in solving (2) to a problem involving Φ_{n+k} and the banded matrix

$$C_{n+k}=(c_{i-j})_{i,j=1}^{n+k} \tag{9}$$

(cf. (7)). The method also entails the solution of a $k\times k$ system. Since there are several algorithms for solving banded Toeplitz systems (see, e.g., [1], [4], [9], [10], [13], and [14]), this procedure should be useful if n is large compared with k. Moreover, we also formulate a procedure that avoids using any of the previously published algorithms for solving banded Toeplitz systems and—as a by-product—provides a new method for this purpose; however, for reasons of stability, this method requires some knowledge of the locations of the zeros of $C(z)$. The method also provides an $O(n)$ procedure for solving (2) when T_n is generated by a rational function. (See § 4.)

2. Derivation of the method. We emphasize that we are not proposing to produce a complete algorithm here. Rather, we are assuming that an algorithm is already available for solving the system (4), where $m = n + p + q = n + k$ henceforth, and we wish to indicate how this can be exploited to solve (2).

Let $\mathscr{F}$ be the underlying field. From (5), (8), and (9),

$$C_m\Phi_m = \begin{bmatrix} [p\times p] & [p\times n] & [p\times q] \\ [n\times p] & T_n & [n\times q] \\ [q\times p] & [q\times n] & [q\times q] \end{bmatrix},$$

where T_n is as in (3) and the other blocks have the indicated dimensions. Therefore, an n-vector X satisfies (2) if and only if

$$C_m\Phi_m \begin{bmatrix} 0_p \\ X \\ 0_q \end{bmatrix} = \begin{bmatrix} U_0 \\ Y \\ V_0 \end{bmatrix},$$

where 0_p and 0_q are zero vectors of dimensions p and q, respectively, $U_0 \in \mathscr{F}^p$, and $V_0 \in \mathscr{F}^q$. For our purposes, it is convenient to view this in the manner stated in the following now obvious lemma.

LEMMA 1. *The system* (2) *has a solution for a given* Y *if and only if there are vectors* U_0 *in* $\mathscr{F}^p$ *and* V_0 *in* $\mathscr{F}^q$ *such that the system*

$$C_m\Phi_m G = \begin{bmatrix} U_0 \\ Y \\ V_0 \end{bmatrix} \tag{10}$$

has a solution G *of the form*

$$G = \begin{bmatrix} 0_p \\ X \\ 0_q \end{bmatrix}, \tag{11}$$

in which case X *satisfies* (2).

Now let $\mathscr{W}$ be the subspace of $\mathscr{F}^m$ consisting of vectors

$$W = [w_{-p+1}, \cdots, w_{n+q}]^t$$

whose components satisfy the homogeneous difference equation

$$\sum_{l=-q}^{p} c_l w_{i-l} = 0, \qquad 1 \leqq i \leqq n, \tag{12}$$

and let

$$W_j = [w_{-p+1}^{(j)}, \cdots, w_{n+q}^{(j)}]^t, \qquad 1 \leqq j \leqq k, \tag{13}$$

form a basis for $\mathscr{W}$. Let

$$F = [f_{-p+1}, \cdots, f_{n+q}]^t \tag{14}$$

be a vector in $\mathscr{F}^m$ whose components satisfy the nonhomogeneous difference equation

$$\sum_{l=-q}^{p} c_l f_{i-l} = y_i, \qquad 1 \leqq i \leqq n. \tag{15}$$

From the definition of C_m, (12) is equivalent to

$$C_m W_j = \begin{bmatrix} U_j \\ 0_n \\ V_j \end{bmatrix}, \qquad 1 \leqq j \leqq k, \tag{16}$$

and (15) is equivalent to

$$C_m F = \begin{bmatrix} U \\ Y \\ V \end{bmatrix}, \tag{17}$$

where U, $U_1, \cdots, U_k$ are in $\mathscr{F}^p$, 0_n is the zero vector in $\mathscr{F}^n$, and V, $V_1, \cdots, V_k$ are in $\mathscr{F}^q$.

There is no doubt about the existence of F and $W_1, \cdots, W_k$; in fact, there are many ways to choose them. We will discuss this in § 3.

THEOREM 1. *Let F and $W_1, \cdots, W_k$ be as just defined. Suppose that for each $j = 1, \cdots, k$ the system*

$$\Phi_m \tilde{W}_j = W_j \tag{18}$$

has a solution

$$\tilde{W}_j = \begin{bmatrix} \tilde{U}_j \\ H_j \\ \tilde{V}_j \end{bmatrix}, \tag{19}$$

and that the system

$$\Phi_m \tilde{F} = F \tag{20}$$

has a solution

$$\tilde{F} = \begin{bmatrix} \tilde{U} \\ \tilde{Y} \\ \tilde{V} \end{bmatrix}, \tag{21}$$

where $\{\tilde{U}, \tilde{U}_1, \cdots, \tilde{U}_k\} \subset \mathscr{F}^p$, $\{\tilde{Y}, H_1, \cdots, H_k\} \subset \mathscr{F}^n$, and $\{\tilde{V}, \tilde{V}_1, \cdots, \tilde{V}_k\} \subset \mathscr{F}^q$. Then the system (2) *has a solution if there are constants $a_1, \cdots, a_k$ such that*

$$\begin{bmatrix} \tilde{U} \\ \tilde{V} \end{bmatrix} = a_1 \begin{bmatrix} \tilde{U}_1 \\ \tilde{V}_1 \end{bmatrix} + \cdots + a_k \begin{bmatrix} \tilde{U}_k \\ \tilde{V}_k \end{bmatrix}, \tag{22}$$

in which case the vector

$$X = \tilde{Y} - a_1 H_1 - \cdots - a_k H_k \tag{23}$$

satisfies (2). *Moreover, the converse is true if Φ_m is invertible.*

Proof. For sufficiency, suppose that (22) holds, and let

$$G = \tilde{F} - a_1 \tilde{W}_1 - \cdots - a_k \tilde{W}_k,$$

which is of the form (11) with X as in (23), from (19), (21), and (22). From (18) and (20),

$$C_m \Phi_m G = C_m (F - a_1 W_1 - \cdots - a_k W_k),$$

and so (16) and (17) imply (10), with

$$\begin{bmatrix} U_0 \\ V_0 \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix} - a_1 \begin{bmatrix} U_1 \\ V_1 \end{bmatrix} - \cdots - a_k \begin{bmatrix} U_k \\ V_k \end{bmatrix}.$$

Therefore, Lemma 1 implies that X as defined by (23) satisfies (2).

For the converse, suppose that Φ_m is invertible and (2) has a solution X. Then the vector G in (11) satisfies (10) for some U_0 in $\mathscr{F}^p$ and V_0 in $\mathscr{F}^q$. From (10) and (17),

$$C_m(F - \Phi_m G) = \begin{bmatrix} U - U_0 \\ 0_n \\ V - V_0 \end{bmatrix},$$

so $F - \Phi_m G \in \mathscr{W}$; hence

$$F - \Phi_m G = a_1 W_1 + \cdots + a_k W_k$$

for some scalars $a_1, \cdots, a_k$. From (18) and (20), this can be rewritten as

$$\Phi_m(\tilde{F} - G) = \Phi_m(a_1 \tilde{W}_1 + \cdots + a_k \tilde{W}_k),$$

so

$$\tilde{F} - G = a_1 \tilde{W}_1 + \cdots + a_k \tilde{W}_k,$$

since Φ_m is invertible. Now (11), (19), and (21) imply (22) and (23). This completes the proof.

THEOREM 2. *Suppose that Φ_m is invertible, let $W_1, \cdots, W_k$ be any basis for $\mathscr{W}$, and let Ψ be the $k \times k$ matrix*

$$\Psi = \begin{bmatrix} \tilde{U}_1 & \cdots & \tilde{U}_k \\ \tilde{V}_1 & \cdots & \tilde{V}_k \end{bmatrix},$$

with $\tilde{U}_1, \cdots, \tilde{U}_k$ and $\tilde{V}_1, \cdots, \tilde{V}_k$ as in (19). *Then T_n is invertible if and only if Ψ is invertible.*

Proof. Since Φ_m is invertible, $\tilde{W}_1, \cdots, \tilde{W}_k$ exist; moreover $\tilde{F}$ exists for every choice of F. If Ψ is invertible, then (22) has a solution $a_1, \cdots, a_k$ for every $\tilde{U}$ and $\tilde{V}$; hence, Theorem 1 implies that (2) has a solution for every Y, and therefore T_n is invertible. For the converse, suppose that Ψ is noninvertible. Then there are constants $b_1, \cdots, b_k$, not all zero, such that

$$b_1 \begin{bmatrix} \tilde{U}_1 \\ \tilde{V}_1 \end{bmatrix} + \cdots + b_k \begin{bmatrix} \tilde{U}_k \\ \tilde{V}_k \end{bmatrix} = \begin{bmatrix} 0_p \\ 0_q \end{bmatrix}.$$

This implies that

$$b_1 \tilde{W}_1 + \cdots + b_k \tilde{W}_k = \begin{bmatrix} 0_p \\ H \\ 0_q \end{bmatrix}, \tag{24}$$

with

$$H = b_1 H_1 + \cdots + b_k H_k$$

(cf. (19)). Because of (18), we can rewrite (24) as

$$b_1W_1+\cdots+b_kW_k=\Phi_m\begin{bmatrix}0_p\\H\\0_q\end{bmatrix}, \tag{25}$$

which makes it apparent that $H\neq 0_n$, since $\{W_1,\cdots,W_k\}$ is linearly independent. Now (16) and (25) imply that

$$C_m\Phi_m\begin{bmatrix}0_p\\H\\0_q\end{bmatrix}=\begin{bmatrix}U_0\\0_n\\V_0\end{bmatrix}, \tag{26}$$

with

$$U_0=\sum_{j=1}^{k}b_jU_j,\qquad V_0=\sum_{j=1}^{k}b_jV_j.$$

However, (26) and Lemma 1 with $Y=0_n$ and $X=H$ imply that $T_nH=0_n$, and therefore T_n is noninvertible, since $H\neq 0_n$.

Henceforth we assume that Φ_m is invertible and that an efficient algorithm is available for solving systems with coefficient matrix Φ_m. Theorem 1 suggests a procedure for solving (2), as follows:

Step 1. Obtain a basis $W_1,\cdots,W_k$ for $\mathcal{W}$, and solve (18) for $\tilde{W}_1,\cdots,\tilde{W}_k$. If (2) is to be solved for more than one Y, then store $\tilde{W}_1,\cdots,\tilde{W}_k$ for repeated use.

Step 2. For the given Y, let F in (14) be a solution of (15), and solve (20) for $\tilde{F}$.

Step 3. Solve the $k\times k$ system (22) for $a_1,\cdots,a_k$. (If (22) has no solution, then (2) has no solution.)

Step 4. Compute X from (23), with $H_1,\cdots,H_k$ as defined in (19) and $\tilde{Y}$ as in (21).

The missing link in this procedure is a discussion of methods for obtaining F and $W_1,\cdots,W_k$. This is the subject of § 3.

3. Computation of F and $W_1,\cdots,W_k$. As mentioned earlier, there are many algorithms specifically designed to solve banded Toeplitz systems efficiently. If C_m is invertible, then we could obtain F by solving (17) with $U=0_p$ and $V=0_q$ by means of one of these algorithms. We could also obtain $W_1,\cdots,W_k$ by solving (16) in this way, with

$$\begin{bmatrix}U_1&\cdots&U_k\\V_1&\cdots&V_k\end{bmatrix}=I_k.$$

However, all algorithms for solving banded Toeplitz systems require some kind of assumption on C_m; in fact, most require that C_m and all its principal submatrices be invertible. Therefore, we will suggest a recursive method for computing suitable vectors F and $W_1,\cdots,W_k$. This method requires no specific assumptions on C_m (even that it be invertible), and it addresses the question of stability; however, it does require information on the zeros of $C(z)$.

One solution (14) of (15) can be obtained from the recursion

$$f_i = \frac{1}{c_{-q}}\left[y_{i-q} - \sum_{l=-q+1}^{p} c_l f_{i-q-l}\right], \qquad q+1 \leqq i \leqq n+q, \tag{27}$$

with $f_i = 0$, if $-p+1 \leqq i \leqq q$. To exhibit a basis for $\mathscr{W}$, we first consider the Maclaurin expansion

$$[z^q C(z)]^{-1} = \sum_{\nu=0}^{\infty} \alpha_\nu z^\nu.$$

The $\{\alpha_\nu\}$ can be computed recursively

$$\alpha_\nu = -\frac{1}{c_{-q}} \sum_{l=-q+1}^{p} c_l \alpha_{\nu-q-l}, \qquad \nu \geqq 1, \tag{28}$$

with $\alpha_\nu = 0$ if $\nu < 0$ and $\alpha_0 = 1/c_{-q}$. The vectors (13) with

$$w_i^{(j)} = \alpha_{i-j+p}, \quad -p+1 \leqq i \leqq n+q, \quad 1 \leqq j \leqq k, \tag{29}$$

form a basis for $\mathscr{W}$. To see that they satisfy (12), observe that if (29) holds, then

$$\sum_{l=-q}^{p} c_l w_{i-l}^{(j)} = \sum_{l=-q}^{p} c_l \alpha_{i-j+p-l}. \tag{30}$$

However, from (28)

$$\sum_{l=-q}^{p} c_l \alpha_{\mu-l} = 0, \qquad \mu > -q.$$

Therefore, the right side of (30) vanishes if $i \geqq 1$ and $1 \leqq j \leqq k$, since then $i-j+p > -q$. To see that $W_1, \cdots, W_k$ are linearly independent, it suffices to observe that the first k rows of the $(n+k) \times k$ matrix

$$[W_1, \cdots, W_k] \tag{31}$$

form an upper triangular matrix with $1/c_{-q}$ in each diagonal position; hence, (31) has rank k.

This procedure provides a formal method for obtaining F and $W_1, \cdots, W_k$; however, it is computationally useless for large n if $C(z)$ has zeros in $|z| < 1$. To be specific, let $z_1, \cdots, z_L$ be the distinct zeros of $C(z)$, with respective multiplicities $m_1, \cdots, m_L$. Then

$$\alpha_i = \sum_{l=1}^{L} p_l(i) z_l^{-i},$$

where p_l is a polynomial of degree $m_l - 1$. This means that the sequence $\{\alpha_i\}$ grows very rapidly with increasing i if $|z_l| < 1$ for one or more values of l. Since the recursion (27) has the explicit solution

$$f_i = \sum_{\nu=1}^{i-q} \alpha_{i-\nu-q} y_\nu, \qquad q+1 \leqq i \leqq n+q,$$

with $f_i = 0$ if $-p+1 \leqq i \leqq q$, f_i also becomes large as i increases. Therefore, these recursions can lead to overflow for large n. Moreover, it is well known that the propagation

of errors renders the recursion formula (27) useless if $|z_l| < 1$ for some l. (To a lesser extent, the presence of repeated roots on $|z| = 1$ is also a source of instability.)

If $C(z)$ has no zeros *outside* the unit circle, then it makes sense to replace the recursion (27) by

$$f_{n-i}=\frac{1}{c_p}\left[y_{n+p-i}-\sum_{l=-q}^{p-1} c_l f_{n+p-i-l}\right], \qquad p\leqq i\leqq n+p-1,$$

with $f_{n-i} = 0$ if $-q \leqq i \leqq p-1$. This also yields a solution (14) of (15). To obtain a basis for $\mathscr{W}$ in this case, we consider the Laurent series

$$[z^{-p}C(z)]^{-1}=\sum_{\nu=0}^{\infty}\beta_\nu z^{-\nu},$$

convergent for large z. The $\{\beta_\nu\}$ can be computed recursively

$$\beta_\nu=-\frac{1}{c_p}\sum_{l=-q}^{p-1} c_l\beta_{\nu-p+l}, \qquad \nu>0, \tag{32}$$

with $\beta_\nu = 0$ if $\nu < 0$ and $\beta_0 = 1/c_p$. The vectors (13) with

$$w_i^{(j)}=\beta_{n-p+j-i}, \quad -p+1\leqq i\leqq n+q, \quad 1\leqq j\leqq k, \tag{33}$$

form a basis for $\mathscr{W}$. To see that they satisfy (12), observe that if (33) holds, then

$$\sum_{l=-q}^{p} c_l w_{i-l}^{(j)}=\sum_{l=-q}^{p} c_l\beta_{n-p+j-i+l}. \tag{34}$$

However, from (32),

$$\sum_{l=-q}^{p} c_l\beta_{\mu+l}=0, \qquad \mu>-p;$$

therefore, the right side of (34) vanishes if $i \leqq n$ and $j \geqq 1$, since then $n - p + j - i > -p$. To see that $W_1, \cdots, W_k$ are linearly independent, observe that in this case the last k rows of (31) form an upper triangular matrix with $1/c_p$ in each diagnoal position.

Since

$$\beta_i=\sum_{l=1}^{L} q_l(i)z_l^i,$$

where q_l is a polynomial of degree $m_l - 1$, it can be shown that this method of computation is stable if $|z_l| \leqq 1$ and $m_l = 1$ if $|z_l| = 1$ $(1 \leqq l \leqq L)$.

If $C(z)$ has zeros in both the interior and exterior of the unit disc, then the recursive procedures that we have considered so far are both unstable. We will now propose a procedure applicable to this situation. It requires that we know a factorization

$$C(z) = z^{s-q}A(z)B(1/z), \tag{35}$$

where

$$A(z)=\sum_{\mu=0}^{r} a_\mu z^\mu \qquad (a_0a_r\neq 0),$$

and

$$B(z) = \sum_{\nu=0}^{s} b_\nu z^\nu \qquad (b_0 b_s \neq 0),$$

with $r > 0$, $s > 0$, and $r + s = p + q = k$, such that $A(z)$ has no zeros in $|z| < 1$, $z^s B(1/z)$ has no zeros in $|z| > 1$, and $A(z)$ and $z^s B(1/z)$ have no zeros in common. (This last assumption is clearly superfluous if $C(z) \neq 0$ for $|z| = 1$; however, if $C(z)$ has zeros on $|z| = 1$, it may be convenient to allocate them between $A(z)$ and $z^s B(1/z)$ subject to this restriction. This would be so, for example, if $C(z) = C(1/z)$, so that C_m is symmetric. In this case an appropriate factorization would be $C(z) = A(z)A(1/z)$, where the zeros of $A(z)$ are in $|z| \leqq 1$.)

Since $A(z)$ and $z^s B(1/z)$ are relatively prime by assumption, there are unique polynomials $g(z)$ and $h(z)$ such that deg $g < r$, deg $h < s$, and

$$z^s g(z) B(1/z) + h(z) A(z) = 1; \tag{36}$$

moreover, the coefficients of $g(z)$ and $h(z)$ can be found by solving a $k \times k$ linear system. Now define

$$Y(z) = \sum_{l=1}^{n} y_l z^l$$

and

$$\tilde{Y}(z) = \sum_{l=1}^{n} y_{n-l+1} z^{l-1},$$

and notice that

$$Y(z) = z^n \tilde{Y}(1/z). \tag{37}$$

Consider the expansions

$$\frac{Y(z)}{A(z)} = \sum_{i=0}^{\infty} \xi_i z^{i+1} \tag{38}$$

and

$$\frac{\tilde{Y}(1/z)}{B(1/z)} = \sum_{i=0}^{\infty} \eta_i z^{-i}. \tag{39}$$

Notice that $\{\xi_i\}$ and $\{\eta_i\}$ can be computed recursively, as follows:

$$\xi_i = \frac{1}{a_0}\left[y_{i+1} - \sum_{l=1}^{r} a_l \xi_{i-l} \right], \qquad i \geqq 0, \tag{40}$$

and

$$\eta_i = \frac{1}{b_0}\left[y_{n-i+1} - \sum_{l=1}^{s} b_l \eta_{i-l} \right], \qquad i \geqq 0, \tag{41}$$

where, for convenience, we define $y_i = 0$ if $i \leqq 0$ or $i \geqq n + 1$, and $\xi_i = \eta_i = 0$ if $i < 0$.

Because of the assumptions on the zeros of $A(z)$ and $B(1/z)$, the recursions (40) and (41) are stable, or at worst, mildly unstable if $C(z)$ has repeated zeros on $|z| = 1$.

Now define the formal Laurent series

$$F(z) = z^{q+1}g(z)\sum_{i=0}^{\infty}\xi_i z^i + z^{n+q-s}h(z)\sum_{i=0}^{\infty}\eta_i z^{-i} = \sum_{l=-\infty}^{\infty} f_l z^l. \tag{42}$$

Then (35), (36), (37), (38), and (39) imply that $C(z)F(z) = Y(z)$. Therefore, (14) with $f_{-p+1}, \cdots, f_{n+q}$ as in (42) satisfies (15).

To obtain a basis $W_1, \cdots, W_k$ for $\mathscr{W}$, we first define

$$\Gamma(z) = z^s g(z)\sum_{\mu=0}^{\infty}\alpha_\mu z^\mu + h(z)\sum_{\mu=0}^{\infty}\beta_\mu z^{-\mu} = \sum_{l=-\infty}^{\infty}\gamma_l z^l, \tag{43}$$

where

$$[A(z)]^{-1} = \sum_{\mu=0}^{\infty}\alpha_\mu z^\mu \tag{44}$$

and

$$[B(1/z)]^{-1} = \sum_{\mu=0}^{\infty}\beta_\mu z^{-\mu}. \tag{45}$$

The coefficients $\{\alpha_\mu\}$ and $\{\beta_\mu\}$ can be computed recursively, with $\alpha_\mu = \beta_\mu = 0$ if $\mu < 0$, $\alpha_0 = 1/a_0$, $\beta_0 = 1/b_0$,

$$\alpha_\mu = -\frac{1}{a_0}\sum_{l=1}^{r} a_l\alpha_{\mu-l}, \qquad \mu \geqq 1,$$

and

$$\beta_\mu = -\frac{1}{b_0}\sum_{l=1}^{s} b_l\beta_{\mu-l}, \qquad \mu \geqq 1.$$

It is shown in [7] (see also [6]) that the Toeplitz matrix

$$\Gamma_{n+k} = (\gamma_{i-j})_{i,j=1}^{n+k}$$

is invertible. We will now show that the first r and last s columns of Γ_{n+k} form a basis for $\mathscr{W}$. (This follows from the main result of [7]; however, we include its brief verification here for the reader's convenience.) To see this, let W be the νth column of Γ_{n+k}, i.e.,

$$W = [w_{-p+1}, \cdots, w_{n+q}]^t = [\gamma_{i-\nu}, \cdots, \gamma_{n+k-\nu}]^t,$$

so that $w_i = \gamma_{i+p-\nu}$, $-p+1 \leqq i \leqq n+q$. Then

$$\sum_{l=-q}^{p} c_l w_{i-l} = \sum_{l=-q}^{p} c_l\gamma_{i-l+p-\nu}, \tag{46}$$

which is the coefficient of $z^{i+p-\nu}$ in the formal Laurent expansion of $C(z)\Gamma(z)$. However, (35), (36), (43), (44), and (45) imply that $C(z)\Gamma(z) = z^{s-q}$; therefore, the right side of (46)

vanishes for $1 \leqq i \leqq n$ provided that $i + p - \nu \neq s - q$ for $1 \leqq i \leqq n$. This condition holds if $1 \leqq \nu \leqq r$ or $n + r + 1 \leqq \nu \leqq n + k$, which proves our assertion.

4. Toeplitz systems with matrices generated by rational functions. If $\Phi(z) = 1$, then T_n in (3) is the $n \times n$ band matrix

$$T_n = (c_{i-j})_{i,j=1}^{n},$$

and $\Phi_m = I_m$. Now Steps 1–4 of § 2 simplify to yield a procedure for solving (2) in which the only simultaneous system to be dealt with is of order k.

Step 1. Obtain $W_1, \cdots, W_k$ recursively, as described in § 3. If (2) is to be solved for more than one Y, store these vectors.

Step 2. Obtain F recursively, as described in § 3.

Step 3. Solve the $k \times k$ system

$$\sum_{j=1}^{k} a_j w_i^{(j)} = f_i, \quad -p+1 \leqq i \leqq 0, \quad n+1 \leqq i \leqq n+q$$

for $a_1, \cdots, a_k$. (If this is impossible, then (2) has no solution.)

Step 4. Compute

$$x_i = f_i - \sum_{j=1}^{k} a_j w_i^{(j)}, \qquad 1 \leqq i \leqq n.$$

The number of operations required for this procedure is $O(kn)$ as n (as compared to $O(k^2 n)$ for methods for solving general $n \times n$ banded systems that do not have the Toeplitz structure). Although there are many "fast" methods for solving banded Toeplitz systems, most of them require recursion with respect to n and are based on the assumption that the principal submatrices of T_n are all invertible. Moreover, to the author's knowledge, the stability of these methods has not been studied, except insofar as Bunch's results [2] on stability of algorithms for general Toeplitz systems apply to them.

In the situation that we have just discussed, the matrices $\{T_n\}$ can be described as being generated by the Laurent polynomial $C(z)$. Now we consider the case where they are generated by the rational functions

$$T(z) = \frac{C(z)}{P(z)Q(1/z)},$$

where $C(z)$ is as in (6),

$$P(z) = \sum_{l=0}^{\mu} p_l z^l$$

and

$$Q(z) = \sum_{l=0}^{\nu} q_l z^l.$$

We assume here that $(\mu + \nu) p_0 q_0 p_\mu q_\nu \neq 0$, and that no two of the polynomials $P(z)$, $Q(1/z)$ and $C(z)$ have a common zero. Here we let $\Phi(z)$ be the formal Laurent expansion of

$$R(z) = [P(z)Q(1/z)]^{-1}$$

obtained as follows:

(a) If $Q = 1$, then

$$R(z) = [P(z)]^{-1} = \sum_{l=0}^{\infty} \phi_l z^l,$$

so that the matrices (5) are lower triangular.

(b) If $P = 1$, then

$$R(z) = [Q(1/z)]^{-1} = \sum_{l=-\infty}^{0} \phi_l z^l,$$

so that the matrices (5) are upper triangular.

(c) If $\mu > 0$ and $\nu > 0$, then $\Phi(z)$ is obtained from $P(z)$ and $Q(z)$ in the same way that $\Gamma(z)$ (cf. (42)) was obtained from $A(z)$ and $B(1/z)$ in § 3. (There is no need to assume here that the zeros of $A(z)$ are confined to $|z| \leqq 1$ while those of $B(1/z)$ are in $|z| \geqq 1$; however, if these conditions hold with strict inequalities, then $\Phi(z)$ is the unique Laurent series which converges to $T(z)$ in an annulus containing $|z| = 1$.)

In this situation, the inverses $\{\Phi_m^{-1}\}$ are banded matrices that are "quasi-Toeplitz" in a sense made explicit in [7], and systems of the form (4) can be solved explicitly with a number of operations that are $O((\mu + \nu)m)$ for large m; moreover, there is no possibility of instability here, since the computation does not involve recursion. Since the formula for Φ_m^{-1} is given in [7], we will not include further detail here. Combining this formula with the recursive methods of § 3 yields the solution of (2) in $O(n)$ (as $n \to \infty$) operations.

In [15] we gave explicit formulas for the solution of (2) when T_n is rationally generated in this way, in terms of Y and determinants involving the values of $P(z)$ and $Q(1/z)$ at the zeros of $C(z)$. Although some discussion of numerical implementation was included in [15], the principal interest there was in the formulas. To the author's knowledge, the only previously published $O(n)$ algorithm specifically designed to solve $n \times n$ systems with rationally generated Toeplitz matrices is due to Dickinson [5]. However, Dickinson's method requires that $T_1, \cdots, T_n$ all be invertible, and he did not consider stability.

REFERENCES

[1] G. BECK, *A fast algorithm for the solution of banded Toeplitz sets of linear equations*, Alkalmaz. Mat. Lapok, 8 (1982), pp. 157–176. (In Hungarian, with English summary.)

[2] J. R. BUNCH, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.

[3] R. P. BRENT, F. G. GUSTAVSON, AND D. Y. Y. YUN, *Fast solution of Toeplitz systems of equations and computation of Padé approximants*, J. Algorithms, 1 (1980), pp. 259–295.

[4] B. W. DICKINSON, *Efficient solution of linear equations with banded Toeplitz matrices*, IEEE Trans. Acoust., Speech, Signal Process., ASSP-27 (1979), pp. 421–423.

[5] ———, *Solution of linear equations with rational Toeplitz matrices*, Math. Comp., 34 (1980), pp. 227–233.

[6] T. N. E. GREVILLE, *On a problem concerning band matrices with Toeplitz inverses*, Proc. 8th Manitoba Conf. Numer. Math. Comput., 1978, pp. 275–283; Utilitas Mathematica Publ. Winnipeg.

[7] T. N. E. GREVILLE AND W. F. TRENCH, *Band matrices with Toeplitz inverses*, Linear Algebra Appl., 27 (1979), pp. 199–209.

[8] F. G. GUSTAVSON AND D. Y. Y. YUN, *Fast algorithms for rational Hermite approximation and solution of Toeplitz systems*, IEEE Trans. Circuits and Systems, CAS-26 (1979), pp. 750–755.

[9] S. B. HALEY, *Solution of band matrix equations by projection-recurrence*, Linear Algebra Appl., 32 (1980), pp. 33–48.

[10] A. K. JAIN, *Fast inversion of banded Toeplitz matrices by circular decompositions*, IEEE Trans. Acoust. Speech, Signal Process., ASSP-26 (1978), pp. 121–126.
[11] T. KAILATH, A. VIEIRA, AND M. MORF, *Inverses of Toeplitz operators, innovations, and orthogonal polynomials*, SIAM Rev., 20 (1978), pp. 106–119.
[12] W. F. TRENCH, *An algorithm for the inversion of finite Toeplitz matrices*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 512–522.
[13] ———, *Inversion of Toeplitz band matrices*, Math. Comp., 28 (1974), pp. 1089–1095.
[14] ———, *Explicit inversion formulas for Toeplitz band matrices*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 167–179.
[15] ———, *Solution of systems with Toeplitz matrices generated by rational functions*, Linear Algebra Appl., 74 (1986), pp. 191–211.
[16] P. WHITTLE, *Prediction and Regulation*, van Nostrand, Princeton, NJ, 1963.
[17] S. ZOHAR, *Toeplitz matrix inversion: The algorithm of W. F. Trench*, J. Assoc. Comput. Mach., 16 (1969), pp. 592–601.
[18] ———, *The solution of a Toeplitz set of linear equations*, J. Assoc. Comput. Mach., 21 (1974), 272–276.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

006

PERMANENTAL INEQUALITIES FOR CORRELATION MATRICES*

ROBERT GRONE† AND STEPHEN PIERCE†‡

Abstract. Let A be a positive semidefinite Hermitian matrix of order n with $|a_{11}| = \cdots = |a_{nn}| = 1$. We prove that per $(A) \geqq (1/n)\|A\|^2$, where $\|A\|$ is the Frobenius norm of A. This follows from a stronger result when $n = 4$, namely per $(A) \geqq \frac{1}{3}(\|A\|^2 - 1)$. Various corollaries are obtained.

Key words. permanent, norm, inequality, correlation matrix

AMS(MOS) subject classifications. 15A15, 15A42

1. Notation and results. An n-by-n matrix A is a *correlation matrix* if and only if A is positive semidefinite Hermitian and $a_{ii} = 1$, for all $i = 1, \cdots, n$. The convex compact set of correlation matrices is denoted by C_n. For $A, B \in C_n$ we will use the notation $A \equiv B$ to denote that A is similar to B by a diagonal unitary matrix and/or a permutation matrix. We will let $\|A\|^2 = \text{trace}\,(A^*A) = \sum_{i,j} |a_{ij}|^2$ denote the Frobenius norm of A and per (A) denote the permanent of A. We remark that if $A \equiv B$, then $\|A\| = \|B\|$ and per $(A) =$ per (B). We let J_n denote the n-by-n matrix all of whose entries equal 1.

When $n = 2$, per $(A) = \frac{1}{2}\|A\|^2$ for all $A \in C_2$. In this paper we establish the following inequality for $n = 3, 4$.

THEOREM 1. *Let $A \in C_n$, $n = 3$ or 4. Then*

$$\text{per}\,(A) \geqq 1 + \frac{1}{3}\sum_{i \neq j} |a_{ij}|^2. \tag{1}$$

Furthermore, equality holds if and only if either: $A = I_3$, $A = I_4$, $A \equiv Y_3 = \frac{3}{2}I_3 - \frac{1}{2}J_3$, or $A \equiv 1 \oplus Y_3$.

Theorem 1 yields the following immediate corollary in the case $n = 3$.

COROLLARY 1. *If $A \in C_3$, then* per $(A) \geqq \frac{1}{3}\|A\|^2$, *with equality if and only if either $A = I_3$ or $A \equiv Y_3$.*

The $n = 4$ case of Theorem 1 is used to obtain an analogue of Corollary 1 when $n \geqq 4$. Suppose $n = 4m + r$, where $0 \leqq r \leqq 3$. Let P denote the set of all partitions of $N = \{1, \cdots, n\}$ into sets $N_1, \cdots, N_m, N_{m+1}$ where $|N_i| = 4$, $i = 1, \cdots, m$, and $|N_{m+1}| = r$. For any subset S of N, let $A[S]$ denote the principal submatrix of A with rows and columns corresponding to S.

THEOREM 2. *If $A \in C_n$, $n \geqq 4$, then*

$$\frac{1}{|P|}\sum_{P}\prod_{i=1}^{m+1} \text{per}\,(A[N_i]) \geqq \frac{1}{n}\|A\|^2. \tag{2}$$

Furthermore, equality holds if and only if $A = I_n$.

Theorem 2 can be used to obtain the following theorem.

THEOREM 3. *If $A \in C_n$, $n \geqq 2$, then*

$$\text{per}\,(A) \geqq \frac{1}{n}\|A\|^2. \tag{3}$$

Furthermore, equality holds if and only if either: $n = 2$, $A = I_n$, or $n = 3$ and $A \equiv Y_3$.

* Received by the editors June 10, 1987; accepted for publication August 21, 1987.

† Department of Mathematical Sciences, San Diego State University, San Diego, California 92182.

‡ The work of this author was partially supported by National Science Foundation grant DMS-8601959.

Theorem 3 implies the following corollaries, each of which was conjectured in [1].

COROLLARY 2. *If* $A \in C_n$, *then*

$$\text{per}(A) \geqq \prod_{i=1}^{n} \left(\sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/n}. \tag{4}$$

Furthermore, equality holds if and only if either: $A = I_n$, $A \equiv J_2$, *or* $A \equiv Y_3$.

We remark that (4) can be restated as

$$\text{per}(A) \geqq (h(A^2))^{1/n}, \qquad A \in C_n,$$

where $h(A) = \prod_{i=1}^{n} a_{ii}$. In this form, Corollary 2 resembles a result in [3], namely

$$(h(A^n))^{1/n} \geqq \text{per}(A) \quad A \text{ positive semidefinite.}$$

Our next result is the following corollary.

COROLLARY 3. *If* $A \in C_n$ *and* $\det(A) = 0$, *then*

$$\text{per}(A) \geqq \frac{n}{n-1}.$$

Furthermore, equality holds if and only if either $A \equiv J_2$ *or* $A \equiv Y_3$.

Corollary 3 can be used to prove our next result.

THEOREM 4. *Suppose* A *is* $n \times n$ *positive semidefinite. Then*

$$(n-1)\,\text{per}(A) + \det(A) \geqq nh(A).$$

Furthermore, equality holds if and only if either: A *is diagonal,* A *has a zero row,* $n = 2$, *or* $A \equiv DY_3D$, *where* D *is a nonnegative diagonal matrix.*

In [4] the second author has conjectured that the minimum value of the permanent for $A \in C_n$ with A singular occurs exactly when $A \equiv Y_n = (n/(n-1))I_n - (1/(n-1))J_n$.

Our last theorem validates this conjecture for $n = 4$ and follows from the techniques involved in the proof of Lemma 3.

THEOREM 5. *Let* $A \in C_4$ *with* $\det(A) = 0$. *Then* $\text{per}(A) \geqq 40/27$ *with equality holding if and only if* $A \equiv Y_4$.

2. Lemmas. For $A \in C_n$, we define the function

$$f(A) = \text{per}(A) - \left(1 + \frac{1}{3} \sum_{i \neq j} |a_{ij}|^2\right).$$

We let $\lambda_n \geqq \lambda_{n-1} \geqq \cdots \geqq \lambda_1 \geqq 0$ denote the eigenvalues of A.

LEMMA 1. *If* $n = 3$ *and* $A \in C_n$, *then*

$$f(A) = g(\lambda) = \tfrac{2}{3}(\lambda_1^2 + \lambda_2^2 + \lambda_3^2) + \lambda_1\lambda_2\lambda_3 - 3.$$

Furthermore, $g(\lambda) \geqq 0$ *with equality if and only if either* $\lambda_1 = \lambda_2 = \lambda_3 = 1$ *or* $\lambda_3 = \lambda_2 = \frac{3}{2} > \lambda_1 = 0$.

LEMMA 2. *Suppose that* $n \geqq 4$, $f(B) \geqq 0$ *for all* $B \in C_{n-1}$, $A \in C_n$ *is positive definite, and* f *has a local minimum in* C_n *at* A. *Then* $f(A) \geqq 0$.

LEMMA 3. *If* $A \in C_4$ *and* $\det(A) = 0$, *then* $f(A) \geqq 0$. *Furthermore, equality holds if and only if* $A \equiv 1 \oplus Y_3$.

3. Proofs.

Proof of Theorem 1. When $n = 3$ Lemma 1 yields (1) and implies that equality holds if and only if $A = I_3$ or $\lambda_3 = \lambda_2 = \frac{3}{2} > \lambda_1 = 0$. In the second case, $\frac{3}{2}$ must be an

eigenvalue of every principal 2-by-2 submatrix of A by Cauchy Interlacing. This implies that $A \equiv Y_3$.

Suppose now that $n = 4$ and that f has a local minimum at A in C_4. If A is positive definite, then Lemma 2 implies that $f(A) \geqq 0$, which amounts to (1). Furthermore, since for $B \in C_3$, $f(B) > 0$ unless $B = I_3$ or $B \equiv Y_3$, an examination of the proof of Lemma 2 will show that $f(A) > 0$ unless $A = I_4$ in the case when A is positive definite. If $A \in C_4$ is singular, then Lemma 3 completes the proof. □

Proof of Theorem 2. If $n = 4$, then the left-hand side of (2) is just per (A) and by Theorem 1

$$\operatorname{per}(A) \geqq 1 + \frac{1}{3}\sum_{i \neq j} |a_{ij}|^2 \geqq \frac{1}{4}\|A\|^2.$$

Hence we may assume $n \geqq 5$.

Also by Theorem 1 we have that for $n \geqq 5$

$$\begin{aligned} \prod_{i=1}^{m} \operatorname{per} A[N_i] &\geqq \prod_{i=1}^{m} \left[1 + \frac{1}{3}(\|A[N_i]\|^2 - 4)\right] \\ &\geqq 1 + \frac{1}{3}\sum_{i=1}^{m} (\|A[N_i]\|^2 - 4). \end{aligned} \tag{5}$$

Furthermore, equality will hold in (5) if and only if $A[N_i] = I_4$, $i = 1, \cdots, m$.

Suppose now that $n = 4m$ or that $n \equiv 0 \pmod 4$. Then, using (5) we observe that the left-hand side of (2) is greater than

$$1 + \frac{1}{3|P|}\sum_{P}\sum_{i=1}^{m} (\|A[N_i]\|^2 - 4). \tag{6}$$

Furthermore, the equality will be strict unless all principal 4-by-4 submatrices of A equal I_4, in which case $A = I_n$. Now consider that if $i \neq j$, then $|a_{ij}|^2$ occurs in the double summation in (6) exactly $|P|(3n/(n^2 - n))$ times. Hence (6) becomes

$$\begin{aligned} 1 + \frac{1}{3|P|}\left(|P|\frac{3n}{n(n-1)}\sum_{i \neq j} |a_{ij}|^2\right) &= 1 + \frac{1}{n-1}\sum_{i \neq j} |a_{ij}|^2 \\ &\geqq 1 + \frac{1}{n}\sum_{i \neq j} |a_{ij}|^2 = \frac{1}{n}\|A\|^2. \end{aligned}$$

This completes the proof when $n \equiv 0 \pmod 4$. Furthermore, since the inequality in (5) is strict unless $A = I_n$, the inequality in (2) is strict unless $A = I_n$.

In the case when $n = 4m + 1$, $|a_{ij}|^2$ occurs in the summation (6) exactly $|P|(3(n-1)/(n^2 - n))$ times, and in this case the expression in (6) equals $(1/n)\|A\|^2$. Hence (2) is valid, and the case of equality is again only when $A = I_n$.

In the case when $n = 4m + 2$, we use

$$\prod_{i=1}^{m+1} \operatorname{per}(A[N_i]) \geqq 1 + \frac{1}{3}\sum_{i=1}^{m} (\|A[N_i]\|^2 - 4) + \frac{1}{2}(\|A[N_{m+1}]\|^2 - 2). \tag{7}$$

Furthermore, equality holds in (7) if and only if $A = I_n$. As before, we sum (7) over P and obtain that the left side of (2) is greater than

$$1 + \frac{1}{3|P|}\sum_{P}\sum_{i=1}^{m} (\|A[N_i]\|^2 - 4) + \frac{1}{2|P|}\sum_{P}(\|A[N_{m+1}]\|^2 - 2). \tag{8}$$

Now we count that $|a_{ij}|^2$ occurs exactly $|P|((3n-6)/(n^2-n))$ times in the first summation of (8) and exactly $|P|(2/(n^2-n))$ times in the second summation of (8). Hence (8) reduces to $(1/n)\|A\|^2$. Furthermore, the inequality in (2) is strict unless equality holds in (7), in which case $A = I_n$.

Finally, suppose that $n = 4m + 3$. In this case the left side of (2) is greater than

$$1+\frac{1}{3|P|}\sum_{P}\left[\left(\sum_{i=1}^{m}\|A[N_i]\|^2-4\right)+\|A[N_{m+1}]\|^2-3\right]. \tag{9}$$

Counting the occurrences of $|a_{ij}|^2$ as before yields that (9) is equal to

$$1+\frac{1}{3|P|}\left(|P|\frac{3(n-3)}{n^2-n}+|P|\frac{6}{n^2-n}\right)(\|A\|^2-n)=\frac{1}{n}\|A\|^2.$$

Hence inequality (2) holds, and, as before, the inequality is strict unless $A = I_n$. □

Proof of Theorem 3. The case $n = 2$ is clear since per $(A) = \frac{1}{2}\|A\|^2$ in this instance. The case $n = 3$ is covered by Corollary 1. The case when $n \geqq 4$ follows from a result of Lieb [2] which implies that each summand on the left-hand side of (2) is dominated by per (A). Hence their average, the left side of (2), is also dominated by per (A). □

Proof of Corollary 2. For $A \in C_n$, let $r_i = \sum_{j=1}^{n}|a_{ij}|^2$. Then

$$\text{per}\,(A)\geqq\frac{r_1+\cdots+r_n}{n}\geqq(r_1\cdots r_n)^{1/n}.$$

The first inequality is from Theorem 3, while the second is the arithmetic-geometric mean inequality. □

Proof of Corollary 3. Suppose $A \in C_n$, det $(A) = 0$, and $\lambda_n \geqq \cdots \geqq \lambda_2 \geqq \lambda_1 = 0$ are the eigenvalues of A. By Theorem 3

$$\text{per}\,(A)\geqq\frac{\lambda_n^2+\cdots+\lambda_2^2}{n}. \tag{10}$$

Since $\lambda_n + \cdots + \lambda_2 = n$, the right-hand side of (10) is minimized when

$$\lambda_n=\cdots=\lambda_2=\frac{n}{n-1}=\frac{\lambda_n^2+\cdots+\lambda_2^2}{n}.$$

Proof of Theorem 4. Suppose A is positive semidefinite. If A is diagonal or has a zero row, then there is nothing to prove. We may also dispense with the cases $n = 2$ or 3. Define

$$\varphi(A)=(n-1)\,\text{per}\,(A)+\det(A)-nh(A),$$

so that Theorem 4 can be restated as $\varphi(A) \geqq 0$. Let $D = \text{diag}\,(a_{11}^{-1/2}, \cdots, a_{nn}^{-1/2})$, and let $A_1 = DAD$. Notice that $\varphi(A_1) = \varphi(A)/h(A)$, so that it will suffice to show that $\varphi(A_1) \geqq 0$. In other words, we may assume that $A \in C_n$. If $A \in C_n$ and A is singular, then Corollary 3 yields that $\varphi(A) \geqq 0$.

If A is positive definite then A can be expressed as $A = \alpha E_{11} + A_0$, where $\alpha > 0$, and A_0 is positive semidefinite and singular. We may assume without loss of generality that $A_0 \in C_n$. Let $A_x = xE_{11} + A_0$ and define $\bar{\varphi}(x) = \varphi(A_x)$, so that

$$\bar{\varphi}(x)=\varphi(A_0)+x[(n-1)\,\text{per}\,A(1)+\det A(1)-n],$$

where $A(1)$ is the principal submatrix of A obtained by deleting row 1 and column 1. By the singular case, we know that $\varphi(A_0) \geqq 0$. If we assume an inductive hypothesis on n and apply it to $A(1)$, we obtain that the coefficient of x in $\bar{\varphi}(x)$ is

$$(n-1)\operatorname{per}A(1)+\det A(1)-n=[\operatorname{per}A(1)-1]+[(n-2)\operatorname{per}A(1)+\det A(1)-(n-1)]>0.$$

Since $\varphi(A) = \bar{\varphi}(\alpha) > 0$, we are finished. □

Proof of Theorem 5. Let the eigenvalues of A be $\lambda_4 \geqq \lambda_3 \geqq \lambda_2 \geqq \lambda_1 = 0$, and let $E_k(\lambda)$ denote the kth elementary symmetric function of $\lambda = (\lambda_4, \lambda_3, \lambda_2, \lambda_1)$ for $k = 1, 2, 3, 4$. The permanent can be expressed as

$$\operatorname{per}(A) = 18 - 4E_2(\lambda) + 2E_3(\lambda) + 2f_1(A)$$

where

$$f_1(A) = |a_{12}a_{34}|^2 + |a_{13}a_{24}|^2 + |a_{14}a_{23}|^2.$$

In the proof of Lemma 3 we establish inequality (17) for singular A in C_4 which is

$$2f_1(A) \geqq 2E_3(\lambda) - 2E_2(\lambda) + 6.$$

From this we see that

$$\operatorname{per}(A) \geqq 24 - 6E_2(\lambda) + 4E_3(\lambda). \tag{11}$$

For a function such as that in the right-hand side of (11) we know that the minimum is achieved when all nonzero λ_i's are equal. Thus we easily verify that the minimum value of the right side of (11) is 40/27 and occurs when $\lambda_4 = \lambda_3 = \lambda_2 = \frac{4}{3}$.

Now suppose that $\operatorname{per}(A) = 40/27$ and $\lambda_4 = \lambda_3 = \lambda_2 = \frac{4}{3} > \lambda_1 = 0$. In this case $A = \frac{4}{3}I_4 - B$ where B is rank 1 Hermitian with $b_{11} = \cdots = b_{44} = \frac{1}{3}$. In this case $B \equiv \frac{1}{3}J_4$ and $A \equiv Y_4$. □

Proof of Lemma 1. The expression for $f(A)$ in terms of $\lambda_1, \lambda_2, \lambda_3$ follows since

$$\operatorname{per}(A) - \det(A) = \sum_{i \neq j} |a_{ij}|^2,$$

$$\det(A) = \lambda_1\lambda_2\lambda_3,$$

and

$$\sum_{i \neq j} |a_{ij}|^2 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 - 3.$$

The inequality and cases of equality follow from calculus. □

Proof of Lemma 2. If $A = 1 \oplus B$, $B \in C_{n-1}$, we are done since $f(A) = f(B) \geqq 0$. Hence we may assume that $|a_{12}|^2 + \cdots + |a_{1n}|^2 > 0$. For $x \geqq 0$, let

$$A_x = \begin{bmatrix} 1 & xa_{12} & \cdots & xa_{1n} \\ xa_{21} & & & \\ \vdots & & & A(1) \\ xa_{n1} & & & \end{bmatrix},$$

and let

$$h(x) = f(A_x) = \operatorname{per}A(1) - 1 - \tfrac{1}{3}(\|A(1)\|^2 - (n-1)) + x^2[\operatorname{per}A - \operatorname{per}A(1) - \tfrac{2}{3}(|a_{12}|^2 + \cdots + |a_{1n}|^2)].$$

Since A is positive definite, so is A_x for all $1-\varepsilon < x < 1+\varepsilon$ and some $\varepsilon > 0$. Since A is a local minimum for f on C_n, we must have that $h'(1) = 0$, or that

$$\operatorname{per} A - \operatorname{per} A(1) - \tfrac{2}{3}(|a_{12}|^2 + \cdots + |a_{1n}|^2) = 0.$$

Now we apply the assumption concerning C_{n-1} to $A(1)$ in the previous expression to obtain

$$\operatorname{per} A - [1 + \tfrac{1}{3}(\|A(1)\|^2 - (n-1))] - \tfrac{2}{3}(|a_{12}|^2 + \cdots + |a_{1n}|^2) \geqq 0$$

or

$$f(A) \geqq 0.$$

Note that if the inequality for $A(1) \in C_{n-1}$ is strict, then $f(A) > 0$. □

Proof of Lemma 3. Suppose $A \in C_4$ and $\det(A) = 0$. Let $E_k(\lambda) = E_k(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ denote the kth elementary symmetric function of the eigenvalues of A for $k = 1, 2, 3, 4$. Let $f_1(A)$ and $f_2(A)$ be the respective sums of the diagonal products of A corresponding to the products of two disjoint transpositions and four cycles. More specifically,

$$f_1(A) = |a_{12}a_{34}|^2 + |a_{13}a_{24}|^2 + |a_{14}a_{23}|^2$$

and

$$f_2(A) = 2\ \mathrm{Re}\ (a_{12}a_{23}a_{34}a_{41} + a_{12}a_{24}a_{31}a_{43} + a_{13}a_{24}a_{32}a_{41}).$$

By the arithmetic-geometric mean inequality it is clear that

$$2f_1(A) \geqq |f_2(A)| \geqq f_2(A). \tag{12}$$

Each of $f(A)$ and $E_k(\lambda)$ can be expressed in terms of the diagonal products of A. Using this and that $E_1(\lambda) = 4$, $E_4(\lambda) = 0$, the following can be seen:

$$f(A) = 13 - \tfrac{10}{3}E_2(\lambda) + 2E_3(\lambda) + 2f_1(A). \tag{13}$$

Evaluating $\det(A - I_4)$ yields

$$\begin{aligned} E_2(\lambda) - E_3(\lambda) - 3 &= -(\lambda_4 - 1)(\lambda_3 - 1)(\lambda_2 - 1) \\ &= -\det(A - I_4) \\ &= f_1(A) - f_2(A) \end{aligned}$$

or

$$f_1(A) = f_2(A) + E_2(\lambda) - E_3(\lambda) - 3. \tag{14}$$

From (12) we see that $f_2(A) \geqq -2f_1(A)$ which together with (14) yields

$$f_1(A) \geqq -2f_1(A) + E_2(\lambda) - E_3(\lambda) - 3$$

or

$$2f_1(A) \geqq \tfrac{2}{3}(E_2(\lambda) - E_3(\lambda) - 3). \tag{15}$$

Using (15) together with (13) yields

$$f(A) \geqq g_1(\lambda) = 11 - \tfrac{8}{3}E_2(\lambda) + \tfrac{4}{3}E_3(\lambda). \tag{16}$$

Using (12) and (14) also shows that

$$f_1(A) \leqq 2f_1(A) - (E_3(\lambda) - E_2(\lambda) + 3)$$

or

$$2f_1(A) \geqq 2E_3(\lambda) - 2E_2(\lambda) + 6. \tag{17}$$

Using (17) together with (13) establishes:

$$f(A) \geqq g_2(\lambda) = 19 - \tfrac{16}{3}E_2(\lambda) + 4E_3(\lambda). \tag{18}$$

We have now established that $f(A)$ is bounded below by each of the two functions of the spectrum of A, $g_1(\lambda)$ and $g_2(\lambda)$. The proof now splits into two cases.

In the first case, suppose the second smallest eigenvalue of A satisfies $0 \leqq \lambda_2 \leqq 1$. In this case $g_1(\lambda) \geqq 0$ when λ is constrained by $\lambda_4 \geqq \lambda_3 \geqq \lambda_2 \geqq 0$, $\lambda_2 \leqq 1$ and $\lambda_4 + \lambda_3 + \lambda_2 = 4$. In fact, under these constraints, $g_1(\lambda) = 0$ if and only if $\lambda_4 = \lambda_3 = \frac{3}{2}$ and $\lambda_2 = 1$.

In the second case, suppose that $1 \leqq \lambda_2 \leqq \frac{4}{3}$. In this case $g_2(\lambda) \geqq 0$ when λ is constrained by $\lambda_4 \geqq \lambda_3 \geqq \lambda_2 \geqq 1$ and $\lambda_4 + \lambda_3 + \lambda_2 = 4$. Furthermore, $g_2(\lambda) = 0$ if and only if $\lambda_4 = \lambda_3 = \frac{3}{2}$ and $\lambda_2 = 1$.

In either case we have that $f(A) \geqq 0$ and that $f(A) = 0$ implies $\lambda_4 = \lambda_3 = \frac{3}{2}$ and $\lambda_2 = 1$. If $f(A) = 0$ we must also have that $f(A) = g_1(\lambda)$ and $f(A) = g_2(\lambda)$. This implies that $f_2(A) = -2f_1(A)$ and $f_2(A) = 2f_1(A)$ so that equality holds in (15) and (17). This yields $f_1(A) = 0$. Using $f_1(A) = 0$ together with $\lambda_4 = \lambda_3 = \frac{3}{2}$ and $\lambda_2 = 1$ yields that $A \equiv 1 \oplus Y_3$. □

4. A conjecture. We conjecture that Theorem 1 holds when $n \geqq 5$, and that equality will hold if and only if either $A = I_n$, or $A \equiv I_{n-3} \oplus Y_3$. Since Lemma 2 is valid for general n, it would suffice to establish that Lemma 3 is valid for general n. In view of the example of $A = I_{n-3} \oplus Y_3$, this conjecture would be the best possible bound of the type

$$\text{per}\,(A) \geqq 1 + k(\|A\|^2 - n), \qquad A \in C_n.$$

Since Lemma 2 reduces the conjecture to the singular case it seems as if Lagrange multipliers could be useful. In fact this does yield some interesting information when A is a local extreme for $f(A)$ subject to $A \in C_n$ and $\det(A) = 0$. For simplicity we will illustrate for real A, though the following holds for general A subject to our assumptions.

Let $x_{ij} = x_{ji}$ be real variables for all $i \neq j$. Define $x_{ii} = 0$ and let $X = [x_{ij}]$. Form the Lagrangian

$$L = f(A + X) - \mu \det(A + X).$$

Set the partial of L with respect to x_{ij} equal to 0 and evaluate at $X = 0$ to obtain

$$2 \text{ per } A(i|j) - \tfrac{4}{3}a_{ij} = (-1)^{i+j} 2\mu \det A(i|j) \quad \text{for all } i \neq j \tag{19}$$

where $A(i|j)$ denotes the $(n-1)$-by-$(n-1)$ submatrix of A obtained by deleting row i and column j. Notice that the conditions (19) are quite stringent.

For a fixed i multiply (19) by a_{ij} and sum over $j \neq i$. From the Laplace expansion theorem for determinants and permanents, we then have

$$\text{per}\,(A) - \text{per}\,A(i|i) - \frac{2}{3}\sum_{j \neq i} a_{ij}^2 = -\mu \det A(i|i). \tag{20}$$

Taking (19) and (20) together, we obtain

$$p\,\text{adj}\,(A) - \tfrac{2}{3}A = \mu\,\text{adj}\,(A) + D, \tag{21}$$

where

$$p\,\text{adj}\,(A) = [\text{per}\,A(i|j)]$$

and

$$D = \text{diag (per } (A) - \|\text{row}_i\|^2).$$

REFERENCES

[1] R. GRONE AND R. MERRIS, *Conjectures on permanents*, Linear and Multilinear Algebra, to appear.
[2] E. H. LIEB, *Proofs of some conjectures on permanents*, J. Math. Mech., 16 (1966), pp. 127–134.
[3] R. MERRIS, *Extensions of the Hadamard determinant theorem*, Israel J. Math., 46 (1983), pp. 301–304.
[4] S. PIERCE, *Permanents of correlation matrices*, in Current Trends in Matrix Theory, R. Grone and F. Uhlig eds., Elsevier, North-Holland, Amsterdam, 1987.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

007

ON THE EVALUATION OF MATRIX FUNCTIONS GIVEN BY POWER SERIES*

HEINRICH BOLZ† AND WILHELM NIETHAMMER†

Abstract. Matrix power series may slowly converge or even diverge if some eigenvalues of the matrix are near the boundary or outside the disk of convergence. In this case it is proposed to apply suitably chosen summability methods to accelerate or generate convergence; special attention is paid to Euler methods. The matrix logarithm appearing in connection with stationary Markov chains is considered as an example.

Key words. matrix functions, summability methods, logarithm of a matrix

AMS(MOS) subject classifications. 65F30, 65D20

1. Introduction. Matrix functions appear in connection with many applications. Often they can be defined by a power series with a finite positive radius of convergence. The series may be slowly convergent in cases where some eigenvalue is near the boundary of the disk of convergence or it may even be divergent although $f(A)$ is well defined. In both cases, summability methods seem to be appropriate tools for the evaluation of $f(A)$. For the Neumann series $f(A) = \sum_{j=0}^{\infty} A^j$, a lot is known about the application of summability methods (see Eiermann, Niethammer, and Varga [3] and Niethammer and Varga [9]). In this paper, we will show how some of those concepts can be applied to more general functions, e.g., $\ln(I - A)$. Since the exponential is an entire function, these methods cannot be directly applied for the computation of $\exp(A)$.

In § 2, the general concept of the application of summability methods to the computation of $f(A)$ is described. Then it is shown how these results specialize for the class of Euler methods. For measuring the efficiency of a method, the asymptotic convergence factor is introduced. In § 4, some remarks are given concerning the algorithm. Finally, an application is described; it deals with the computation of the matrix logarithm in connection with stationary Markov chains.

2. Summability methods applied to matrix functions. Let f be given by its power series expansion

$$f(z) = \sum_{j=0}^{\infty} u_j z^j, \qquad |z| < R, \tag{2.1}$$

which has a positive radius of convergence R, and let $\mathbf{P} = (\pi_{j,m})_{j \geq 0, 0 \leq m \leq j}$ be an infinite lower triangular matrix with complex entries. A matrix summability method, induced by $\mathbf{P}$, is given by (see [9] for more details)

$$f(z) \sim \sum_{j=0}^{\infty} v_j(z), \quad \text{where} \quad v_j(z) = \sum_{m=0}^{j} \pi_{j,m} u_m z^m. \tag{2.2}$$

Usually, we identify $\mathbf{P}$ and the summability method induced by $\mathbf{P}$.

DEFINITION. $\mathbf{P}$ sums f in a domain $S \subset \mathbb{C}$ if and only if equality holds in (2.2) for all $z \in S$ and, in addition, the convergence is uniform with respect to z in each compact subset of S.

Next, we formulate a well-known result.

* Received by the editors November 20, 1986; accepted for publication July 23, 1987.

† Institut für Praktische Mathematik, Universität Karlsruhe, Englerstrasse 2, D-7500 Karlsruhe, Federal Republic of Germany.

THEOREM 1 (Perron–Okada; see [11]). *Let* **P** *sum the geometric series* $\sum_{j=0}^{\infty} z^j$ *in a domain* $S \subset \mathbb{C}$, *with* $0 \in S$, *and let* f *be holomorphic in a domain* $H \subset \mathbb{C}$. *Then* **P** *sums* f *in*

$$F := \bigcap_{\xi \in \mathbb{C}\setminus H} \xi S = \bigcap_{\xi \in \mathbb{C}\setminus H} \{z \in \mathbb{C}: z/\xi \in S\}. \tag{2.3}$$

Example 1. Consider

$$\ln(1-z) = -\sum_{m=1}^{\infty} \frac{z^m}{m}, \qquad |z| < 1, \tag{2.4}$$

which is holomorphic, e.g., in the domain $H := \mathbb{C}\setminus[1, \infty)$. Then F consists of all $z \in S$ such that $\xi z \in S$ with $0 \leqq \xi \leqq 1$. Thus if S is starlike with respect to 0 and $0 \in S$ then, for the series (2.4), F coincides with S.

In the following, let A denote an n-by-n matrix with complex entries. Usually, $f(A)$ is defined via the Jordan canonical form of A (see [5]).

Now, if f is given by the power series (2.1) and if the spectrum $\sigma(A)$ of A is contained in $D_R(0)$, where $D_\rho(\mu) := \{z \in \mathbb{C}: |z - \mu| < \rho\}$, then $f(A)$ is also given by

$$f(A) = \sum_{j=0}^{\infty} u_j A^j, \tag{2.5}$$

i.e., we have replaced the complex variable z by A in (2.1) [6, Thm. 11.2.3]. Analogously, we conclude the following result from the uniform convergence of the transformed series in (2.2) (see [5, p. 102]).

THEOREM 2. *Let* $\mathbf{P} := (\pi_{j,m})_{j \geqq 0, 0 \leqq m \leqq j}$ *induce a summability method* (2.2), *let* $\sigma(A) \subset F$ *and let* **P** *sum* f *in* F. *Then*

$$f(A) = \sum_{j=0}^{\infty} v_j(A) \quad \textit{where} \quad v_j(A) = \sum_{m=0}^{j} \pi_{j,m} u_m A^m. \tag{2.6}$$

Our aim is to accelerate the convergence of the series (2.5) or—if (2.5) is divergent—to produce a convergent series by the application of a suitable summability method.

Now, some questions arise:

(a) Given **P**, how can a region S be constructed such that $g(z) = 1/(1 - z)$ is summed in S? (Some information about S is necessary for the application of Theorem 1.)

(b) For judging the efficiency of a given method **P**, a measure for the rate of convergence is needed.

(c) Given a function f and a matrix $A \in \mathbb{C}^{n \times n}$, how can **P** be chosen appropriately, i.e., so that the corresponding rate of convergence becomes maximal?

(d) How should the corresponding computation be organized?

For one special matrix function, all these questions have been intensively dealt with, namely for

$$(I - T)^{-1} = \sum_{j=0}^{\infty} T^j \quad \text{if} \quad \rho(T) < 1, \tag{2.7}$$

the Neumann series (see [9]) that appears in connection with the iterative solution of a system of linear equations $x = Tx + c$. A second well-known matrix function is the exponential $\exp(A)$. But since $\exp(z)$ is an entire function the methods proposed here have to be modified; this will be done in a forthcoming paper.

A further important function is

$$\ln(I-T) = -\sum_{m=1}^{\infty} \frac{T^j}{m} \quad \text{if} \quad \rho(T) < 1. \tag{2.8}$$

We will concentrate our considerations on this special function (see [1]). Many results can be applied to other functions defined by a power series with a finite radius of convergence.

In the following sections, we shall examine a special class of summability methods, the so-called Euler methods, because these methods do have—as we shall see—favourable properties concerning the analytic part of our problem as well as its algorithmic solution.

3. Euler methods. Euler methods are used as a tool for numerical analytic continuation (see [7], [8]) and for constructing and analyzing iterative methods for linear systems of equations (see [9]). Here we apply Euler methods to approximate matrix functions.

Euler methods are generated by an Euler function $p(z)$, i.e., $p(0) = 0$, $p(1) = 1$, and p is holomorphic in a neighbourhood of the unit disc D_1 (we write $D_r = D_r(0) = \{z \in \mathbb{C}: |z| < r\}$ and by $\bar{D}_r$ we denote the closure of D_r). Let the power series of p^m be given by

$$(p(z))^m = \sum_{j=m}^{\infty} \pi_{j,m} z^j, \qquad m = 0, 1, 2, \cdots. \tag{3.1}$$

Then, the coefficients $\pi_{j,m,(j\geqq 0, 0\leqq j\leqq m)}$ yield the summability matrix $\mathbf{P}$ of the corresponding Euler method.

The region of summability of the geometric series is given by

$$S(p) = \bar{\mathbb{C}} \backslash \tilde{p}(\bar{D}_1), \tag{3.2}$$

where $\mathbb{C} = \bar{\mathbb{C}} \cup \{\infty\}$ and $\tilde{p} := 1/p$ (see [9]). Thus, we have the information that is necessary for the application of Theorem 1.

As a measure for the rate of convergence of a series $\sum_{m=0}^{\infty} a_m$, we introduce the (asymptotic) convergence factor

$$\kappa = \limsup_{m \to \infty} |a_m|^{1/m}.$$

The region

$$S_r = \{z: \kappa(z) \leqq 1/r\},$$

where for a given summability method the convergence factor of the transformed geometric series summability method is bounded by $1/r$, is

$$S_r = S_r(p) = \bar{\mathbb{C}} \backslash \tilde{p}(D_r), \qquad r > 1. \tag{3.3}$$

By a close inspection of the proof of Theorem 1, we can establish the following result concerning these so-called r-regions of summability (see [11]).

THEOREM 3. *Under the hypotheses of Theorem* 1, *let* $0 \in S_r$ *for some* $r > 1$. *If*

$$z \in F_r = \bigcap_{\xi \in \mathbb{C} \backslash H} \xi S_r = \bigcap_{\xi \in \mathbb{C} \backslash H} \{z \in \mathbb{C}: z/\xi \in S_r\}$$

we have $\kappa(z) \leqq 1/r$ *for the transformed series of the function* f.

The r-region of summability for the logarithm $\ln(1 - z)$ consists of all z for which the line segment $[0, z]$ is contained in S_r; S_r and F_r coincide if S_r is starlike with respect to $z = 0$. We write $F_r(p)$ in the case of an Euler method.

Since we are interested in approximating matrix functions, we need a result concerning the convergence factor of a transformed matrix series. We can guarantee a certain convergence factor if the spectrum of the matrix is contained in an r-region of summability for some function f. This is stated in the next theorem formulated for Euler methods.

THEOREM 4 (see [3]). *Let $\sigma(A) \subseteq F_r(p)$ for an Euler function p and some $r > 1$. Then we have (for an arbitrary matrix norm)*

$$\limsup_{N\to\infty} \left\| f(A) - \sum_{j=0}^{N} v_j(A) \right\|^{1/N} \leqq 1/r \quad \text{with} \quad v_j(A) = \sum_{m=0}^{j} \pi_{j,m} u_m A^m. \tag{3.4}$$

In the proof the Jordan canonical form of A is used. Thus, (3.4) has only an asymptotic meaning, i.e., for finite approximations the rate of convergence may be influenced by the size of the different Jordan blocks (i.e., by the geometric multiplicities of the eigenvalues) and the condition number of the matrix that transforms A to its Jordan form.

Finally, given f and $A \in \mathbb{C}^{n\times n}$, we consider the question of how to choose an appropriate Euler method for the computation of $f(A)$. We assume that we know a compact set $\tilde{T}$ containing the spectrum $\sigma(A)$ and a (large) domain H such that f is holomorphic in H. If now T is defined by

$$T = \{z/\zeta : z \in \tilde{T}, \zeta \in \mathbb{C}\setminus H\} \tag{3.5}$$

then, by Theorems 1 and 2, our problem is reduced to the task of summing the geometric function $g(z) = 1/(1 - z)$ in T by an appropriate Euler method. This last question has been extensively studied, e.g., in [9]. Thus, an Euler method induced by p_0 is "optimal with respect to T" if $S_r(p_0) = T$ for some $r > 1$. An important consequence is the fact that each Euler method induced by p is optimal with respect to all the nonempty sets $S_r(p)$, $r > 1$.

4. Computational remarks. Given $A \in \mathbb{C}^{n\times n}$, a compact set $\tilde{T} \supset \sigma(A)$, f and a domain $H \subset \mathbb{C}$ on which f is holomorphic, we can determine the set T according to (3.5) (remember that for $f(z) = \ln(1 - z)$ and $H = \mathbb{C}\setminus[1, \infty)$, we get $T = \tilde{T}$ if $0 \in \tilde{T}$ and $\tilde{T}$ is starlike with respect to 0). Then we have to determine an optimal Euler method with respect to T, or, at least, find an Euler method such that $T \subset S_r(p)$ for some $r > 1$. Here, it is advantageous to consider the special class of Euler methods generated by

$$p(z) = \frac{s_0 z}{1 - s_1 z - \cdots - s_k z^k}, \qquad k \in \mathbb{N}, \quad s_0 + s_1 + \cdots + s_k = 1. \tag{4.1}$$

For $k = 1$, the sets $S_r(p)$ are disks, for $k = 2$, ellipses and intervals (as degenerated ellipses) (see [8] or [9]). Thus, if T is a disk, an ellipse, or an interval, optimal Euler methods are known. In the general case, it may be easier to enclose T by a disk, etc., rather than to seek the optimal Euler method, because an Euler method of type (4.1) is well suited for computations.

Given p as a power series or in the form (4.1), we have to compute the partial sums

$$q_m(A) = \sum_{j=0}^{m} v_j(A) \quad \text{where} \quad v_j(A) = \sum_{l=0}^{j} \pi_{j,l} u_l A^l, \tag{4.2}$$

of the transformed series for some suitably chosen m. In the general case, the elements $\pi_{j,l}$ of the summability matrix P have to be computed row-wise from the identity $p^m = p \cdot p^{m-1}$. This can be avoided in the special case of the Neumann series, i.e., if $u_m = 1$ ($m = 0, 1, \cdots$); here, recursive formulas for the polynomials q_m can be directly

derived from the coefficients of the power series, respectively, from the parameters $s_0, \cdots, s_k$ of (4.1). For general f and for a p of the form (4.1), a triangular scheme for the computation of $q_m(z)$ is given in [8].

An important point in connection with the computation of matrix functions via (4.2) is that the coefficients of q_m can be calculated by scalar computations. Then, in the last step, $q_m(A)$ has to be evaluated by Horner's scheme. Since no transformation or inversion of a matrix is necessary, this computation can be easily done in a parallel way. By a suitable rearrangement to the operations of the Horner scheme, the number of matrix multiplications needed can be decreased from $m - 1$ to $\sim\sqrt{m}$ (see [12]).

Remark. For a given matrix A, the information used by the algorithm is that the eigenvalues of A are contained in the set $\tilde{T}$. Thus the same polynomial $q_m(A)$ results, regardless of the multiplicities of the eigenvalues. Since it is well known that numerical difficulties sometimes appear in evaluating matrix polynomials in case of multiple eigenvalues (see [6, p. 387]), similar problems have to be expected here.

5. An application to the matrix logarithm. We consider a stationary Markov chain $\{X_t: t \geqq 0\}$ with continuous time and n states. This chain is represented by the semigroup of stochastic matrices $P(t) = (p_{i,j}(t))_{i,j=1}^n$, $t \geqq 0$, of the transition probabilities (see [4])

$$p_{i,j}(t) = p(X_{\tau+t} = j | X_\tau = i), \qquad i, j = 1, \cdots, n, \quad t, \tau \geqq 0.$$

The semigroup is called standard if

$$\lim_{t \to 0} P(t) = P(0) = I.$$

It is well known that, in this case, there exists a matrix $Q = (q_{i,j})_{i,j=1}^n$ with

$$Q = P'(0)$$

(where the right-hand derivative at $t = 0$ is taken elementwise), and we have

$$q_{i,i} \leqq 0, \quad q_{i,j} \geqq 0, \quad i \neq j, \quad i, j = 1, \cdots, n, \quad \sum_{j=1}^{n} q_{i,j} = 0, \quad i = 1, \cdots, n. \tag{5.1}$$

Any matrix Q with property (5.1) is the (right-hand) derivative of a standard stochastic semigroup $\{P(t)\}$ at $t = 0$. Q determines $P(t)$ by the differential equation

$$P'(t) = QP(t), \qquad P(0) = I, \tag{5.2}$$

or

$$P(t) = \exp(Qt). \tag{5.3}$$

We consider the so-called identification problem of Markov chains, i.e., the reconstruction of Q from given data $P(t_0)$ for some $t_0 > 0$ (see [10]). From (5.3), this can be done by evaluating the matrix logarithm

$$Q = \frac{1}{t_0} \ln P(t_0). \tag{5.4}$$

$\ln P(t_0)$ exists if $P(t_0)$ is nonsingular. There appear two questions.

(a) Can $P(t_0)$ be embedded in some Markov chain?

(b) Is this chain (or the stochastic semigroup) uniquely determined? If this is the case, we say the semigroup has a unique logarithm at $t = t_0$.

There is a positive answer to (a) if Q has property (5.1). The second question is answered by the following theorem.

THEOREM 5 (see [2]). *Let* $\{P(t): t \geqq 0\}$ *be a standard stochastic semigroup. If*

$$t_0 \in \hat{T} := \left\{ t > 0 : g(t) = \min_{i=1}^{n} p_{ii}(t) > \frac{1}{2} \right\}, \tag{5.5}$$

then $P(t_0)$ *has a unique logarithm at* $t = t_0$. *Furthermore,* $\hat{T} = (0, \hat{t})$ *for some* $\hat{t} > 0$.

In the following, we consider the case (5.5). Thus, existence and uniqueness are guaranteed. We remark that our methods work not only in this restricted case but also in a more general setting. But if (5.5) is violated, we are not sure—due to possible non-uniqueness—whether we have found the correct answer Q. This reflects the fact that the problem itself might be ill-posed.

To evaluate

$$Q = \frac{1}{t_0} \ln P(t_0),$$

we write

$$\ln P(t_0) = \ln (I - B)$$

with

$$B = I - P(t_0).$$

We need some information about $\sigma(B)$. Let μ be an eigenvalue of B. Then by Gershgorin's Theorem (see [6]), we have

$$\mu \in \bigcup_{i=1} K_i$$

with

$$K_i := \left\{ z \in \mathbb{C} : |z - (1 - p_{i,i}(t_0))| \leqq \sum_{\substack{j=1 \\ j \neq i}}^{n} p_{i,j}(t_0) = 1 - p_{i,i}(t_0) \right\}.$$

Thus by (5.5), it follows that

$$K_i \subseteq \bar{D}(1 - g(t), 1 - g(t)), \qquad i = 1, \cdots, n,$$

and

$$\sigma(B) \subseteq \bar{D}(1 - g(t), 1 - g(t)) \subseteq D(0, 1). \tag{5.6}$$

We see that $\rho(B) < 1$, and thus the series

$$\ln P(t_0) = \ln (I - B) = -\sum_{m=1}^{\infty} \frac{1}{m} B^m \tag{5.7}$$

is convergent to $t_0 Q$.

The problem can be solved approximately by the evaluation of (5.7) but the convergence may be rather slow. The convergence factor is given by

$$k_a := \max_{i=1}^{n} |\mu_i|, \quad \mu_i = 1 - \lambda_i, \quad \lambda_i \in \sigma(P(t_0)), \quad i = 1, \cdots, n$$

(compare the straight line in Fig. 1).

For accelerating the convergence, we want to apply Euler methods of the type (4.1) with $k = 1$ and $k = 2$. For $k = 1$, the Euler function of (4.1) is usually written in the

form $p(z) = \alpha z/(1 - (1 - \alpha)z)$; $S(p)$ is a disk with center $1 - 1/\alpha$ and radius $1/|\alpha|$, $S_r(p)$ is a disk with the same center and radius $1/r|\alpha|$, $r > 1$.

For $k = 2$ the set $S(p)$ is the interior of an ellipse with foci determined by the parameters s_0, s_1, s_2; $S_r(p)$ is a confocal ellipse where the interval between the foci can be seen as a degenerated ellipse (see [8]). The optimal Euler method for the interval $T = [0, z_0]$ belongs to this class of methods; the corresponding parameters are

$$s_0 = 4\theta/z_0, \quad s_1 = -2\theta, \quad s_2 = -\theta^2 \quad \text{with} \quad \theta = (1 - \sqrt{1 - z_0})^2/z_0; \tag{5.8}$$

the asymptotic convergence factor is $|\theta|$.

Thus, we know the regions where the geometric series is summed. But, as long as 0 is contained in the disk, respectively, ellipses, according to our remark at the beginning of § 4 the function $\ln(1 - z)$ is summed in these regions, too.

Thus with respect to the information (5.6), namely that $\sigma(B)$ is contained in $T = \bar{D}(\mu, \mu)$, $0 < 2\mu < 1$, the optimal α for the Euler method with $k = 1$ is given by $\alpha = 1/(1 - \mu)$ with the convergence factor $K_b := \mu/(1 - \mu)$.

If $P(t_0)$ has real eigenvalues (e.g., in the symmetric case) and the information is $\sigma(B) \subseteq [0, \hat{b}]$, $\hat{b} < 1$, we apply the Euler method with $k = 2$ and parameters (5.8) for the interval $T = [0, \hat{b}]$, and we obtain the convergence factor $K_c = |\theta|$.

In Fig. 1 we have plotted the error curves

$$\|\ln(I - B) - f_N(B)\|_\infty$$

for the different approximants $f_N(B)$ which we announced (straight line defines series; dotted line defines Euler method for $k = 1$ with disc information; dashed line defines Euler method for $k = 2$ with interval information).

As an example consider

$$P(t) = \begin{bmatrix} 1+2x & 1-x & 1-x \\ 1-x & 1+2x & 1-x \\ 1-x & 1-x & 1+2x \end{bmatrix}, \qquad x = e^{-1.5t}.$$

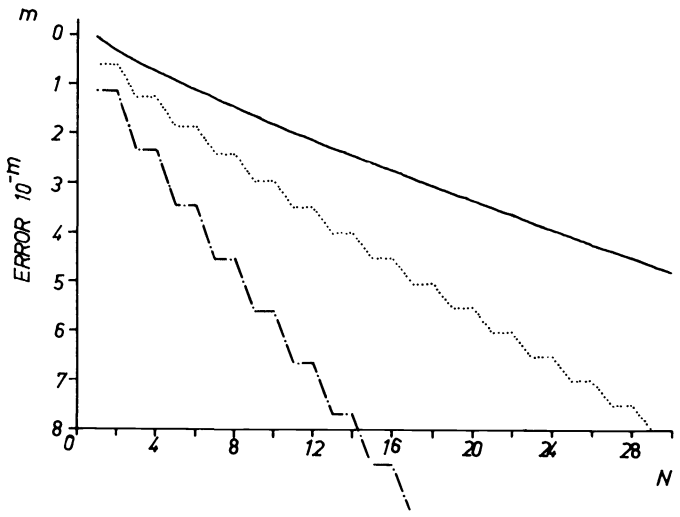


FIG. 1

Then $P(t)$ has eigenvalues 1, x, x, and $P(t) = \exp(Qt)$ with

$$Q = \begin{bmatrix} -1 & 0.5 & 0.5 \\ 0.5 & -1 & 0.5 \\ 0.5 & 0.5 & -1 \end{bmatrix}.$$

Perform the computation for $t_0 = 0.9$, so that $g(t_0) = \min_{i=1}^{3} p_{i,i}(t_0) = 0.5062 > \frac{1}{2}$. The resulting convergence factors are

k_a	k_b	k_c
0.74	0.58	0.32

The decay of the error curves shows the acceleration of convergence by using the Euler–Knopp method enclosing the spectrum $\{0, 1 - x\}$ of B in a circle and the further acceleration by enclosing the spectrum by $[0, 1 - x]$. We remark that the method works also in the case where the original series is divergent, i.e., if we have a problem of analytic continuation rather than a problem of convergence acceleration.

REFERENCES

[1] H. BOLZ, *Methoden zur Auswertung von Matrixfunktionen insbesondere des Logarithmus*, Fakultät für Mathematik, Universität Karlsruhe. 1986 (Dr.rer.nat.).

[2] J. R. CUTHBERT, *On the uniqueness of the logarithm for Markov semi-groups*, J. London Math. Soc., 4 (1972), pp. 623–630.

[3] M. EIERMANN, W. NIETHAMMER, AND R. S. VARGA, *A study of semiiterative methods for nonsymmetric systems of linear equations*, Numer. Math., 47 (1985), pp. 505–533.

[4] D. FREEDMAN, *Markov Chains*, Holden-Day, San Francisco, CA, 1971.

[5] F. R. GANTMACHER, *The Theory of Matrices, Vol.* 1, Chelsea, New York, 1960.

[6] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[7] W. NIETHAMMER, *Ein numerisches Verfahren zur analytischen Fortsetzung*, Numer. Math., 21 (1972), pp. 81–92.

[8] ———, *Numerical applications of Euler's series transformation and its different generalizations*, Numer. Math., 34 (1980), pp. 271–283.

[9] W. NIETHAMMER AND R. S. VARGA, *The analysis of k-step iterative methods for linear systems from summability theory*, Numer. Math., 41 (1983), pp. 177–206.

[10] B. SINGER AND S. SPILERMAN, *The representation of social processes by Markov models*, Amer. J. Sociology, 82 (1976), pp. 1–54.

[11] R. TRAUTNER AND W. GAWRONSKI, *Verschärfung eines Satzes von Borel–Okada über Summierbarkeit von Potenzreiheng*, Period. Math. Hungar., 7 (1976), pp. 201–211.

[12] C. F. VAN LOAN, *A note on the evaluation of matrix polynomials*, IEEE Trans. Automat. Control, AC-24 (1978), pp. 320–321.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

008

AN ALGORITHM FOR SUBSPACE COMPUTATION, WITH APPLICATIONS IN SIGNAL PROCESSING*

DANIEL R. FUHRMANN†

Abstract. An algorithm for computing the eigenvectors corresponding to the m algebraically smallest or largest eigenvalues of an $n \times n$ symmetric matrix $\mathbf{A}$ is described. The algorithm consists of repeated applications of the Rayleigh–Ritz procedure to a sequence of subspaces of dimension $m + 1$ which converges to the desired subspace. The method is closely related to the Lanczos method, but requires a constant amount of computation at each iteration. Applications of the algorithm include the adaptive covariance eigenstructure computation, in which the matrix $\mathbf{A}$ can change while the algorithm is in progress.

Key words. subspace, eigenvectors, eigenvalues, Lanczos algorithm, Rayleigh–Ritz procedure, signal processing

AMS(MOS) subject classifications. 65, 94

1. Introduction.

An area of much recent activity in signal processing is the use of eigenvector and singular value decompositions (SVDs) in extracting information from time series or sensor array data. Figuring prominently in this work are Pisarenko Harmonic Retrieval [13] and the Tufts–Kumaresan SVD method [16] for line spectrum estimation, Schmidt's MUSIC algorithm for bearing estimation [14], and other applications of the "signal subspace" concept in beamforming [15].

In all of the above problems, the dominant computational requirement is the calculation of a set of eigenvectors corresponding to the m algebraically smallest (or largest) eigenvalues of an $n \times n$ symmetric matrix $\mathbf{A}$, where $m \ll n$. For applications in which n is large and $\mathbf{A}$ is structured, the algorithm of choice is the Lanczos algorithm [7], [12]. A variant of the Lanczos method which exhibits faster convergence in certain applications has also been proposed by Davidson [3]. Both the Lanczos and Davidson methods have the property that the computation and storage requirements grow at each iteration. Variations of the Lanczos method that require constant computation and storage have also been proposed; these are the s-step method of Karush [11] and the block s-step method of Cullum and Donath [2]. The distinguishing characteristic of all of these algorithms is that they are based on the forward matrix-vector multiplication $\mathbf{Ax}$, which poses no numerical difficulties, and which can often be implemented with far less than n^2 multiplications.

In this paper we propose a Lanczos-like algorithm which is similar in spirit to the s-step methods. The algorithm consists of repeated applications of the Rayleigh–Ritz (RR) procedure [9], [12] to a sequence of subspaces which converges to the desired invariant subspace. The advantages of the new method are that the RR portion of the algorithm is highly structured in a way that may lead to significant computational savings, and that it requires little storage beyond that of $\mathbf{A}$ and the vectors that span the desired subspace. In addition, all of the computations lend themselves well to parallel implementation.

The application of an algorithm such as this in signal processing goes beyond the use of the algorithm's final output. The method for going from one iterate to the next

* Received by the editors March 12, 1987; accepted for publication October 1, 1987. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† Electronic Systems and Signals Research Laboratory, Department of Electrical Engineering, Washington University, St. Louis, Missouri 63130.

can in fact form a stochastic approximation method when the data itself is time-varying. A well-known example of this is Widrow's least-mean-squares (LMS) algorithm, in which the gradient search technique, an optimization method for deterministic problems, is modified for a statistical estimation problem. Recent research [6], [17] has indicated that the same concept can be applied to eigenvector decomposition of sample covariance matrices.

The paper is organized as follows. Section 2 contains the basic algorithm description, along with a brief discussion of its convergence properties. Section 3 discusses the relationship of this method to the Lanczos and other methods. Sections 4 and 5 discuss some of the computational issues associated with the algorithm. Section 6 describes an example, and § 7 discusses the applications of the method in signal processing and in computing the SVD.

2. Algorithm description. The following notation shall hold throughout. $\mathbf{A}$ is an $n \times n$ symmetric matrix, with eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ in ascending order, and corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$. $\mathbf{x}_1 \cdots \mathbf{x}_m$ are m orthonormal vectors which satisfy $\mathbf{x}_i^T\mathbf{A}\mathbf{x}_j = d_i\delta_{ij}$, $\mathbf{D} = \operatorname{diag}(d_1 \cdots d_m)$, $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]$, and $X = \operatorname{span}\{\mathbf{x}_1 \cdots \mathbf{x}_m\}$. $\mathbf{Z}$ is an $(m+1) \times (m+1)$ orthonormal matrix to be described shortly; $\mathbf{Z}^-$ is $\mathbf{Z}$ with the last column removed. The superscript (k) refers to the value of the base quantity at the kth iteration of the algorithm.

Formally, the method is as follows:

For $k = 1, 2, \cdots$ to convergence, do:

1. Given $\mathbf{X}$, $\mathbf{D}$
2. Compute $\mathbf{A}\mathbf{x}_m$
3. $\mathbf{y} = \mathbf{A}\mathbf{x}_m - d_m\mathbf{x}_m$
4. $\mathbf{y} = \mathbf{y}/|\mathbf{y}|$
5. $\mathbf{X}^+ = [\mathbf{X}|\mathbf{y}]$
6. $\mathbf{H} = (\mathbf{X}^+)^T\mathbf{A}\mathbf{X}^+$
7. Eigenvector decomposition of $\mathbf{H}$: $\mathbf{H} = \mathbf{Z}^T\mathbf{D}\mathbf{Z}$
8. $\mathbf{X} = \mathbf{X}^+\mathbf{Z}^-$

At iteration k the subspace $X^{(k)}$ is augmented by the vector $\mathbf{y}$. $\mathbf{y}$ is orthogonal to $X^{(k)}$ by the condition on the $\mathbf{x}_i$ given above. The augmented subspace is termed X^+.

Steps 6–8 define the RR procedure applied to X^+. The returned vectors $\mathbf{x}_i^{(k+1)}$ are vectors in X^+ closest to eigenvectors of $\mathbf{A}$, in the sense that they are stationary points of the Rayleigh quotient evaluated over this subspace. The m vectors $\mathbf{x}_1^{(k+1)} \cdots \mathbf{x}_m^{(k+1)}$ corresponding to the smaller Ritz values $d_1 \cdots d_m$ form the basis for $X^{(k+1)}$.

The $(m+1) \times (m+1)$ matrix $\mathbf{H}$ has a form that makes it simple to compute. Since $\mathbf{x}_i^T\mathbf{A}\mathbf{x}_j = d_i\delta_{ij}$, by the previous RR step, it follows that only the terms $\mathbf{y}^T\mathbf{A}\mathbf{x}_i$ need be computed. $\mathbf{H}$ then has the form

$$\mathbf{H} = \begin{bmatrix} d_1 & & & & b_1 \\ & d_2 & & & b_2 \\ & & \cdot & & \cdot \\ & & & \cdot & b_m \\ b_1 & b_2 & \cdot & b_m & d_{m+1} \end{bmatrix} \tag{2.1}$$

where $b_i = \mathbf{y}^T\mathbf{A}\mathbf{x}_i$.

This form of $\mathbf{H}$ lends insight into the convergence properties of the algorithm. By the Cauchy Interlace Theorem, $d_i^{(k+1)} \leqq d_i^{(k)}$ with equality if and only if $b_i^{(k)} = 0$, or equivalently, if and only if $\mathbf{x}_i$ is an eigenvector. Furthermore, $d_i^{(k)}$ is bounded below by

TABLE 2.1

	No. of multiplications
Computation of **Ax**	p
Computation of normalized **y**	$3n$
Computation of **H**	$p + mn$
Eigendecomposition of **H**	$O(m^3)$
Backtransformation	nm^2
Total	$2p + (m^2 + m + 3)n + O(m^3)$

λ_i by the Courant–Fischer minimax characterization of the eigenvalues. Since d_i is strictly decreasing and bounded below, it must converge; we hypothesize that it does indeed converge to λ_i. (This has been proven true for $m = 1$ in [6].) It follows that $\mathbf{x}_i$ must converge directionally to the eigenvector associated with λ_i. In our initial simulations it is observed that each of the vectors $\mathbf{x}_2 \cdots \mathbf{x}_m$ converge in turn to eigenvectors as they become deficient in the eigenvectors that preceded them.

The number of multiplications required by the algorithm as described are given in Table 2.1. The number of multiplications in the computation of **Ax**, p, depends on the structure of **A** and the nature of the multiplication algorithm. In the worst case, $p = n^2$, but in many practical situations, say where **A** is sparse, $p \ll n^2$. Typically the computation of **Ax** is thought of as a single subroutine call.

3. Relationship to other methods. As indicated in § 1, our method is closely related to the Lanczos method and its variants.

The Lanczos method generates a sequence of subspaces, termed Krylov subspaces, defined by

$$K^{(j)} = \text{span}\,\{\mathbf{x}_0, \mathbf{A}\mathbf{x}_0, \cdots, \mathbf{A}^{j-1}\mathbf{x}_0\} \tag{3.1}$$

where $\mathbf{x}_0$ is some initial vector. For every j, an orthogonal basis for $K^{(j)}$ is generated such that the **H** matrix for the RR procedure is a tridiagonal matrix; furthermore, **H** is computed recursively so that only two new elements are added to this tridiagonal matrix at each iteration. Note that the $K^{(j)}$ form a sequence of subspaces of increasing size.

The s-step method of Karush [11], discussed in Faddeev and Faddeeva [5], starts with an initial vector $\mathbf{x}_0$ and performs s iterations of the Lanczos algorithm. Then the RR procedure is implemented on the Krylov subspace $K^{(s)}(\mathbf{x}_0)$. The smallest (or largest) Ritz vector is used as the new initial vector, and this process is repeated. The block s-step method of Cullum and Donath [2] takes an orthogonal set of m vectors $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m]$ and computes the subspace $(\mathbf{X}\ \mathbf{A}\mathbf{X} \cdots \mathbf{A}^{(s-1)}\mathbf{X})$, then chooses the best m Ritz vectors to form the next **X**. Both methods require a constant amount of storage and computation at each iteration.

The present method generates only one new vector at each iteration, and the best m Ritz vectors from the $m + 1$-dimensional subspace are saved. This leads to a smaller requirement in storage (although this is probably not an important issue) and to the structured **H** matrix whose eigendecomposition forms the heart of the RR procedure. If the dimension of $X^{(k)}$ were allowed to grow with each k, then our method would be equivalent to the Lanczos method in terms of the sequence of subspaces generated.

The present method can be thought of as a generalization of our "rotational methods" [6] which are equivalent to the present method for $m = 1$ and $m = 2$. The $m = 1$ case is also equivalent to s-step Lanczos with $s = 1$, first considered by Hestenes and Karush [10].

4. Eigendecomposition of H. The form of **H** leads to an eigendecomposition algorithm based on the iterative solution of the characteristic equation. It is straightforward to show that

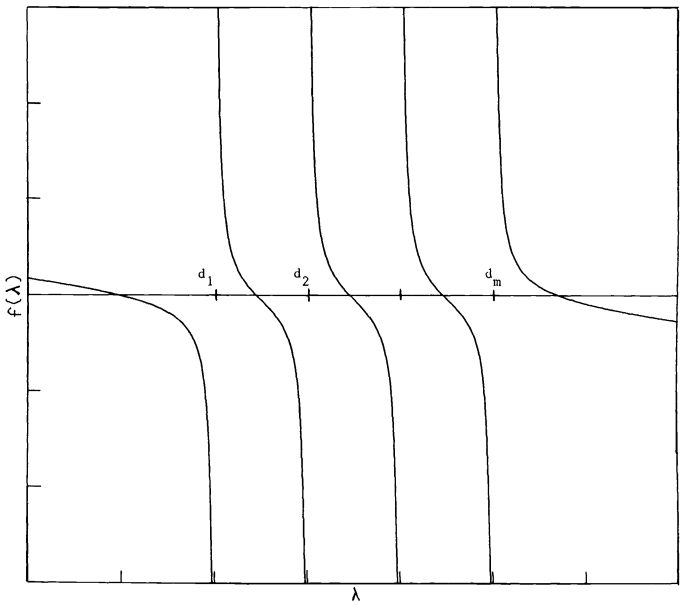
$$\det(\mathbf{H}-\lambda\mathbf{I}) = \left[\prod_{i=1}^{m}(d_i-\lambda)\right]\left[(d_{m+1}-\lambda)-\sum_{j=1}^{m}\frac{b_j^2}{(d_j-\lambda)}\right]. \tag{4.1}$$

Solution of the characteristic equation can be accomplished by solving for the roots of the second factor. The function

$$f(\lambda) = (d_{m+1}-\lambda)-\sum_{j=1}^{m}\frac{b_j^2}{(d_j-\lambda)} \tag{4.2}$$

offers a simple proof of the Cauchy Interlace property. $f(\lambda)$ has a negative derivative everywhere except the singularities at $\lambda = d_j$. As can be seen in Fig. 4.1, the roots of $f(\lambda)$ are separated by the d_i.

Equation (4.2) is similar to Golub's "secular equation" [8] which arises in the solution of the eigensystem of a diagonal-plus-rank-1 matrix. The solution of the secular equation plays a central role in Dongarra and Sorenson's parallel algorithm for the complete symmetric eigenvalue problem [4]. In our method, **H** is diagonal-plus-rank-2. A straightforward approach to solving for the roots of $f(\lambda)$ such as bisection or Newton's method exhibits difficulties when there are repeated d_is or very small b_is. Unfortunately, these are the exact conditions one expects in our signal processing applications or as the algorithm converges. Cheng [1] has reported an algorithm for computing this eigendecomposition which overcomes these difficulties and exhibits quadratic convergence to each of the Ritz values. Since each of the Ritz values is well localized, it is conceivable that their computation could be carried out independently and in parallel.



SECULAR EQUATION

FIG. 4.1

5. Loss of orthogonality. The basic algorithm described in § 2 suffers from the same difficulty as the unmodified Lanczos algorithm, namely the loss of orthogonality of the basis vectors $\mathbf{x}_1 \cdots \mathbf{x}_m$. This loss of orthogonality comes about through roundoff errors in either the computation of $\mathbf{y}$ or the back-transformation $\mathbf{X}^{(k+1)} = \mathbf{X}^{+}\mathbf{Z}^{-}$, and worsens as the algorithm progresses.

A remedy for this problem would be to replace steps 3–4 of the algorithm with a complete Gram–Schmidt (GS) orthogonalization of the set of vectors

$$\{\mathbf{x}_1, \cdots, \mathbf{x}_m, \mathbf{A}\mathbf{x}_m - d_m\mathbf{x}_m\}$$

at every iteration. Although this solution is inelegant, it does not drastically alter the computational requirements, since one GS orthogonalization requires the same order of magnitude of computation as the backtransformation step ($2nm^2$) and $m \ll n$. This approach is unsatisfactory for the Lanczos algorithm, since m is increasing at each iteration, and (assuming no RR step) there are no other $O(m^2)$ operations in the algorithm.

6. Example. The convergence properties of the algorithm are invariant to orthogonal similarity transformations of the matrix $\mathbf{A}$. As a result, the convergence properties can be studied through analysis or simulations applied to diagonal matrices. In this case the eigenvalues and eigenvectors are known a priori and one can observe the progress of d_i and $\mathbf{x}_i$ toward their known limit points.

For an initial simulation we choose a 100×100 diagonal matrix $\mathbf{A}$ with diagonal elements 0, 1, 2, 3, $\cdots$, 99. m was chosen to be 10.

The algorithm is "primed" by starting with an initial $\mathbf{x}_0$ and letting the dimension of $X^{(k)}$ grow from 1 to 10 in the first 10 iterations. This is equivalent to performing 10 iterations of the Lanczos algorithm, or choosing the Krylov subspace $K^{(m)}(\mathbf{x}_0)$ as the initial subspace $X^{(0)}$. We took

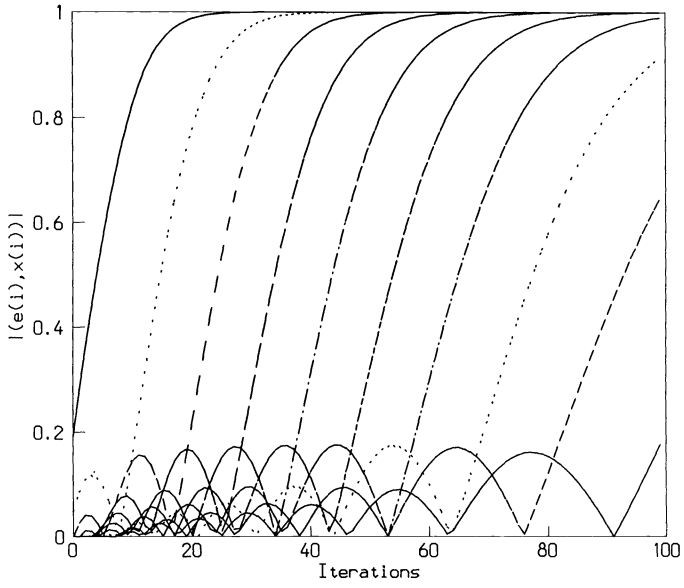
$$\mathbf{x}_0 = \frac{1}{10}[1\ 1\ 1\ \cdots\ 1]^T.$$

$\mathbf{x}_0$ has equal components in all the eigenvectors of $\mathbf{A}$.

Three quantities were measured at each iteration: (1) the inner product of $\mathbf{x}_i$ with the true eigenvector $\mathbf{e}_i$, and (2) the norm of the projection of $\mathbf{x}_i$ onto the minimum invariant subspace span $\{\mathbf{e}_i \cdots \mathbf{e}_m\}$, and (3) the Ritz values d_i. These quantities were plotted against the iteration number; the results are shown in Figs. 6.1–6.3. Figure 6.1 shows how each $\mathbf{x}_i$ converges in turn to the true eigenvector $\mathbf{e}_i$. Figure 6.2 demonstrates that while $\mathbf{x}_i$ may not be close to $\mathbf{e}_i$, the error lies primarily in the desired invariant subspace. Figure 6.3 is in agreement with Fig. 6.1, showing that the smaller Ritz values converge to the true small eigenvalues quickly, with each of the larger Ritz values following in turn.

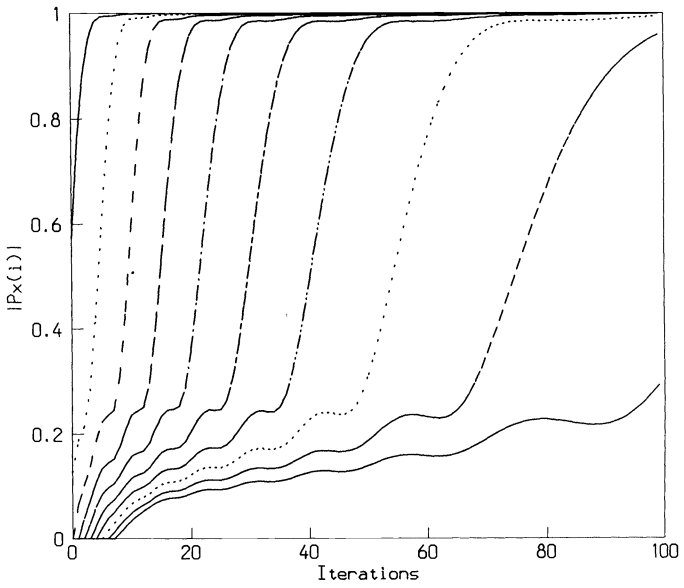
7. Application to signal processing. The present method was conceived as an intermediate step in the problem of adaptively computing invariant subspaces of sample covariance matrices. The goal is an adaptive formulation of Schmidt's MUSIC algorithm [14] or, alternatively, an adaptive eigenvector beamformer, in which the weight vector for an array of antennas or sensors is determined from the eigendecomposition of the received signal covariance matrix. A preliminary report of the application of our method to the MUSIC algorithm can be found in [7].

Two aspects of the present method make it attractive for this problem: (1) the computation is constant at each iteration, and (2) $\mathbf{X}^{(k+1)}$ depends only on $\mathbf{A}$ and $\mathbf{X}^{(k)}$. The first point is important for the design of hardware that must work in parallel with the real-time acquisition of uniformly sampled data. The second point implies that it



EIGENVECTOR CONVERGENCE

FIG. 6.1



SUBSPACE CONVERGENCE

FIG. 6.2

may be possible to change **A** while the algorithm is in progress. If **A**(k) is a converging sequence of covariance estimates, then $X^{(k)}$ should converge to the minimum invariant subspace of **A**(∞). This concept has been successfully applied to Pisarenko Harmonic Retrieval [5], [17], in which the minimum eigenvector of the covariance matrix of a stationary time series is computed.

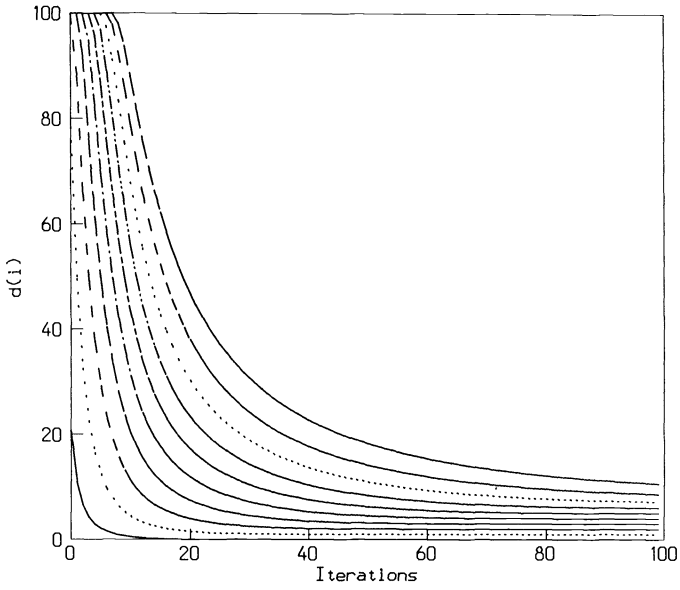


FIG. 6.3

A preferable alternative to computing the eigenstructure of a sample covariance matrix is computing the SVD of the data matrix. The present method, and in fact any eigenvector algorithm, can be applied to this problem. This results from the fact that

$$\mathbf{C} = \begin{bmatrix} 0 & \mathbf{A} \\ \mathbf{A}^T & 0 \end{bmatrix} \tag{7.1}$$

has eigenvalues $\pm\sigma_i$ and 0, and eigenvectors $\begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix}$, where the σ_i are singular values, and $\mathbf{u}_i$ and $\mathbf{v}_i$ are left and right singular vectors, respectively, of $\mathbf{A}$. The application of the standard Lanczos algorithm to SVD computation is discussed in [9].

The complete SVD of a rectangular data matrix $\mathbf{A}$ could be computed using the present method on the matrix $\mathbf{C}$, with the dimension m of the desired subspace equal to the short dimension of $\mathbf{A}$. Each multiplication of the form $\mathbf{Cx}$ could be partitioned into two multiplications $\mathbf{Av}$ and $\mathbf{A}^T\mathbf{u}$, each of which requires mn real multiplications. If fewer singular vectors are desired, one could choose a smaller value of m. If $\mathbf{A}$ were composed of overlapping samples from a time series, such as in the Tufts–Kumaresan method of spectrum estimation, $\mathbf{Av}$ and $\mathbf{A}^T\mathbf{u}$ could be computed via fast convolutions.

8. Summary. We have proposed a numerical method for computing the minimum or maximum invariant subspace of dimension m from a symmetric matrix of order n. It is closely related to the Lanczos method, but requires a constant amount of computation ($O(nm^2) + O(m^3)$) at each iteration. Initial simulations indicate that the small eigenvectors converge rapidly, with each successive vector following in turn. Convergence properties and modification to enhance computational speed are still being studied. Finally, the application of this method to adaptive covariance eigenstructure computation, eigenvector beamforming, and SVD computation is anticipated.

REFERENCES

[1] Y. H. CHENG, *Numerical methods in array processing*, M.S. thesis, Washington University, St. Louis, MO, August 1987.

[2] J. CULLUM AND W. DONATH, *A block Lanczos algorithm for computing the Q algebraically largest eigenvalues and a corresponding eigenspace of large sparse real symmetric matrices*, Proc. 1974 IEEE Conf. on Decision and Control, Phoenix, AZ, pp. 505–509.

[3] E. DAVIDSON, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of a large real-symmetric matrix*, J. Comput. Phys., 17 (1975), pp. 87–94.

[4] J. DONGARRA AND D. SORENSON, *A fully parallel algorithm for the symmetric eigenvalue problem*, Argonne National Laboratory Mathematics and Computer Sciences Division Technical Memorandum No. 62, January 1986.

[5] D. FADDEEV AND N. FADDEEVA, *Computational Methods of Linear Algebra*, W. H. Freeman, San Francisco, CA, 1963.

[6] D. FUHRMANN AND B. LIU, *Rotational search methods for adaptive Pisarenko Harmonic Retrieval*, IEEE Trans. Acoust. Speech Signal Process., December 1986, Vol. ASSP-34, pp. 1550–1565.

[7] D. FUHRMANN, *Adaptive* MUSIC, Proc. Society of Photo-optical Instrumentation Engineers, vol. 826, San Diego, CA, August 1987.

[8] G. GOLUB, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–334.

[9] G. GOLUB AND C. VAN LOAN, *Matrix Computations,* Johns Hopkins University Press, Baltimore, MD, 1983.

[10] M. HESTENES AND W. KARUSH, *A method of gradients for the calculation of the characteristic roots and vectors of a real symmetric matrix*, J. Res. Nat. Bur. Standards, Sec. B, 47 (1951), pp. 471–478.

[11] W. KARUSH, *An iterative method for finding characteristic vectors of a symmetric matrix*, Pacific J. Math., 1 (1951), pp. 233–248.

[12] B. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[13] V. PISARENKO, *The retrieval of harmonics from a covariance function*, Geophysical J. Royal Astronomical Soc., 33 (1973), pp. 347–366.

[14] R. SCHMIDT, *A signal subspace approach to multiple emitter location and spectral estimation*, Ph.D. dissertation, Stanford University, Palo Alto, CA, November 1981.

[15] J. SPEISER, *Progress in eigenvector beamforming*, Proc. SPIE, vol. 564, Real-Time Signal Processing VIII, August 1985.

[16] D. TUFTS AND R. KUMARESAN, *Estimation of frequencies of multiple sinusoids: making linear prediction perform like maximum likelihood*, Proc. IEEE, 70 (1982), pp. 975–989.

[17] R. VACCARO, *On adaptive implementation's of Pisarenko Harmonic Retrieval*, Proc. ICASSP 84, March 1984, p. 6.1.1.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

009

A SYMPLECTIC ORTHOGONAL METHOD FOR SINGLE INPUT OR SINGLE OUTPUT DISCRETE TIME OPTIMAL QUADRATIC CONTROL PROBLEMS*

VOLKER MEHRMANN†

Abstract. A new, numerically stable, structure preserving method for the discrete linear quadratic control problem with single input or single output is introduced, which is similar to Byers' method in the continuous case and faster than the general QZ-algorithm approach of Pappas, Laub, and Sandell.

Key words. symplectic matrices, eigenvalues, invariant subspaces, discrete algebraic Riccati equations, linear quadratic control

AMS(MOS) subject classifications. 65F15, 65H10, 93D15

0. Introduction.

Consider the following discrete optimal control problem:

Minimize

$$J(x_k, u_k) = \sum_{k=0}^{\infty} (y_k^* Q y_k + u_k^* R u_k + x_k^* S u_k + u_k^* S^* x_k)$$

(0.1) subject to the difference equation

$$E x_{k+1} = A x_k + B u_k, \qquad k = 0, 1, 2, \cdots, \qquad x_0 = x^0, \tag{0.2}$$

$$y_k = C x_k, \tag{0.3}$$

where $A, E \in \mathbb{C}^{n,n}$, $Q \in \mathbb{C}^{p,p}$, $C = \mathbb{C}^{p,n}$, $R \in \mathbb{C}^{m,m}$, $B, S \in \mathbb{C}^{n,m}$, and $Q = Q^*$, $R = R^*$ positive semidefinite. Extensive literature has been published in recent years on this subject, in particular concerning numerical algorithms for this problem. See, for example, Arnold [1], Bender and Laub [3], Pappas, Laub, and Sandell [18], Van Dooren [24], Byers [5].

It is well known (e.g., Sage [19]) that this problem has a unique solution if the matrix

$$\begin{bmatrix} C^*QC & S \\ S^* & R \end{bmatrix} =: \mathcal{Q} \tag{0.4}$$

is positive semidefinite and the system of difference equations

$$\begin{bmatrix} A & 0 & B \\ C^*QC & -E^* & S \\ S^* & 0 & R \end{bmatrix} \begin{bmatrix} x_k \\ \mu_k \\ u_k \end{bmatrix} = \begin{bmatrix} E & 0 & 0 \\ 0 & -A^* & 0 \\ 0 & -B^* & 0 \end{bmatrix} \begin{bmatrix} x_{k+1} \\ \mu_{k+1} \\ u_{k+1} \end{bmatrix}, \qquad \begin{matrix} x_0 = x^0 \\ \lim_{k \to \infty} \mu_k = 0 \end{matrix} \tag{0.5}$$

has a unique solution. The system (0.5) has a unique solution if and only if the corresponding matrix pencil

$$\tilde{\mathcal{A}} - \lambda \tilde{\mathcal{B}} := \begin{bmatrix} A & 0 & B \\ C^*QC & -E^* & S \\ S^* & 0 & R \end{bmatrix} - \lambda \begin{bmatrix} E & 0 & 0 \\ 0 & -A^* & 0 \\ 0 & -B^* & 0 \end{bmatrix} \tag{0.6}$$

is a regular pencil, i.e., $\det(\mathcal{A} - \lambda \mathcal{B}) \not\equiv 0$. (See Campbell [12].) The regularity of $\tilde{\mathcal{A}} - \lambda \tilde{\mathcal{B}}$ is guaranteed under the usual system theoretic assumptions, (E, A, B) stabilizable

* Received by the editors May 14, 1987; accepted for publication October 1, 1987. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† Fakultät für Mathematik, Universität Bielefeld, Postfach 8640, 4800 Bielefeld 1, Federal Republic of Germany.

and (E, A, C^*QC) detectable. (The triple (E, A, B) is called stabilizable if the following holds: if $x \in \mathbb{C}^n \setminus \{0\}$, $\lambda \in \mathbb{C}$ and $|\lambda| = 1$ such that $x^*[-\lambda E + A] = 0$, then $x^*B \neq 0$. The triple (E, A, C^*QC) is detectable if the triple (E^*, A^*, C^*QC) is stabilizable.) The condition that $\mathcal{Q}$ is positive semidefinite is also a typical assumption. In many practical problems we even have that Q, R are positive definite and S is zero, so that $\mathcal{Q}$ is positive definite.

In the following we assume that R is positive definite, which means that no costfree controls u_k exist. Then we may reduce system (0.5) and obtain the pencil

$$\begin{aligned} \mathcal{A} - \lambda\mathcal{B} &:= \begin{bmatrix} F & 0 \\ H & E^* \end{bmatrix} - \lambda \begin{bmatrix} I & -G \\ 0 & F^* \end{bmatrix} \\ &:= \begin{bmatrix} A - BR^{-1}S^* & 0 \\ C^*QC - SR^{-1}S^* & E^* \end{bmatrix} - \lambda \begin{bmatrix} E & -BR^{-1}B^* \\ 0 & A^* - SR^{-1}B^* \end{bmatrix} \end{aligned} \tag{0.7}$$

by using

$$u_k = -R^{-1}(S^*x_k + B^*\mu_{k+1}). \tag{0.8}$$

If $\tilde{\mathcal{A}} - \lambda\tilde{\mathcal{B}}$ is regular, then so is $\mathcal{A} - \lambda\mathcal{B}$. Furthermore, since R, Q are positive semidefinite, it follows that H, G are positive semidefinite. It has been shown by Bender and Laub [3] that if E is singular and a solution to the problem (0.1), (0.2), (0.3) exists, which is the case under certain further assumptions (e.g., Bender and Laub [3]), then the pencil (0.7) can be reduced further to a smaller, similar-looking pencil with E nonsingular, using a singular value decomposition of E. Thus, we may assume that E is nonsingular and since it simplifies the description of our algorithm significantly we transform the system such that $E = I$.

Under the assumptions (E, A, B) stabilizable, (E, A, C^*QC) detectable, $E = I$, it is well known (e.g., Pappas, Laub, and Sandell [18]) that the optimal feedback control is

$$u_k = -(R + B^*XB)^{-1}(A^*XB + S)^*x_k, \tag{0.9}$$

where X is the symmetric positive semidefinite solution of the discrete algebraic Riccati equation

$$X = F^*(I + XG)^{-1}XF + H. \tag{0.10}$$

(See also Arnold [1].)

The positive semidefinite solution of (0.10), however, can be obtained via the computation of the deflating subspace, corresponding to the eigenvalues of modulus less than one, of the pencil $\mathcal{A} - \lambda\mathcal{B}$. Let $\begin{bmatrix} Y \\ Z \end{bmatrix}$ be an n-dimensional deflating subspace of $\mathcal{A} - \lambda\mathcal{B}$, i.e.,

$$\mathcal{A}\begin{bmatrix} Y \\ Z \end{bmatrix} = \mathcal{B}\begin{bmatrix} Y \\ Z \end{bmatrix} U, \tag{0.11}$$

where $U \in \mathbb{C}^{n,n}$ has only eigenvalues of modulus less than one. Let Y be invertible, then $X = -ZY^{-1}$ is the positive definite solution of (0.10). Keeping this reduction in mind we now restrict ourselves to the problem of computing the required deflating subspace of $\mathcal{A} - \lambda\mathcal{B}$. (Existence is guaranteed under the above assumptions (e.g., Arnold [1].)

A typical approach to the numerical solution to this problem is the use of the QZ-algorithm of Moler and Stewart [16] as proposed by Pappas, Laub, and Sandell [18]. Unfortunately the QZ-algorithm does not make any use of the symmetries in $\mathcal{A} - \lambda\mathcal{B}$. So it is natural to ask whether a stable QZ-type algorithm can be constructed that takes advantage of the symmetries in $\mathcal{A} - \lambda\mathcal{B}$.

In the continuous time optimal control problem (e.g., Laub [15]), which leads to the pencil

$$\begin{bmatrix} F & G \\ H & -F^* \end{bmatrix} - \lambda \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}, \tag{0.12}$$

this question was posed in a paper by Paige and van Loan [17], and answered in the special case of single output problems, i.e., the matrix H is of rank 1 and $S = 0$, by Byers [4], [5], who also proposed the transfer of his method to the discrete case as a research problem. In this paper we will show that it is possible to construct a similar type algorithm in the discrete case for single input or single output systems (i.e., where rank $H = 1$, $S = 0$ or rank $G = 1$). Considering the pencil $\mathscr{A} - \lambda\mathscr{B}$, it is clear that we may restrict ourselves, without loss of generality, to rank $H = 1$, since if rank $G = 1$ and rank $H \neq 1$, we may permute the pencil and replace λ by $1/\mu$ and have a new pencil $\mathscr{A}' - \mu\mathscr{B}'$ of the same form, which has a rank 1 matrix in the position of H.

1. Notation. By $\mathbb{C}^{n,m}(\mathbb{R}^{n,m})$ we denote the complex (real) $n \times m$ matrices. $\mathbb{C}^{n,1} =: \mathbb{C}^n$, $\mathbb{R}^{n,1} =: \mathbb{R}^n$. Let

$$J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}$$

where I_n is the $n \times n$ identity matrix. (The index n is usually omitted.) By $\sigma(A)$ $(\sigma(A, B))$ we denote the spectrum of $A(A - \lambda B)$; e_i denotes the ith unit vector.

DEFINITION 1.1. A matrix $S \in \mathbb{C}^{2n,2n}$ is called *symplectic* if $SJS^* = J$ (here $*$ denotes the transpose and conjugate). A pencil $A - \lambda B$, $A, B \in \mathbb{C}^{2n,2n}$ is called *symplectic pencil* if $AJA^* = BJB^*$.

Remark 1.2. The usual definition of a symplectic matrix S is $SJS^T = J$, even in the complex case. A matrix S satisfying $SJS^* = J$ is usually called conjugate symplectic. But for simplification, and in order to avoid a permanent distinction between the real and complex case, we use the chosen definition.

DEFINITION 1.3. A matrix $U \in \mathbb{C}^{n,n}$ is called *unitary* if $UU^* = I$.

The set of all symplectic matrices in $\mathbb{C}^{2n,2n}$ is denoted by $\underline{S_{2n}}$, the set of all unitary matrices in $\mathbb{C}^{n,n}$ by $\underline{U_n}$ and we set $\underline{US_{2n}} = U_{2n} \cap S_{2n}$. Observe that U_{2n}, S_{2n}, US_{2n} form multiplicative subgroups of the general linear group. US_{2n} can be characterized as follows (e.g., Paige and van Loan [17]):

$$US_{2n} = \left\{ \begin{bmatrix} Q_1 & Q_2 \\ -Q_2 & Q_1 \end{bmatrix}, Q_1, Q_2 \in \mathbb{C}^{n,n}, Q_1 Q_1^* + Q_2 Q_2^* = I, Q_1 Q_2^* = Q_2 Q_1^* \right\}.$$

In order to perform QR-type algorithms, we need to eliminate certain elements in a matrix using transformations with elementary matrices. Except for the usual Householder transformations and Givens rotations in U_{2n} (e.g., Golub and van Loan [13]), we use the following two types of matrices in US_{2n}:

(i) Let $v \in \mathbb{C}^n$ and $P = I - 2vv^*/v^*v$. Then $H = \left[\begin{smallmatrix} P & 0 \\ 0 & P \end{smallmatrix}\right] \in US_{2n}$ is called a *Householder symplectic* matrix.

(ii) Let $s, c \in \mathbb{C}$, $|s|^2 + |c|^2 = 1$, $s\bar{c} \in \mathbb{R}$,

$$C = \operatorname{diag}(\underbrace{1, \cdots, 1}_{k-1}, c, 1, \cdots, 1), S = \operatorname{diag}(\underbrace{0, \cdots, 0}_{k-1}, s, 0, \cdots, 0);$$

then $J(k, c, s) = \left[\begin{smallmatrix} C & S \\ -S & C \end{smallmatrix}\right] \in US_{2n}$ is called a *Jacobi symplectic* matrix.

As in the usual QR-algorithm we need something like a Hessenberg form and an upper triangular form. For an analysis of useful invariant forms see Bunse-Gerstner [7], [8]. Here we use the following definition.

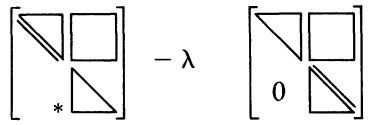
DEFINITION 1.4. (i) A pencil

$$A-\lambda B=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}-\lambda\begin{bmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{bmatrix}$$

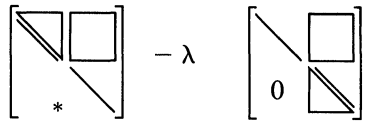
with A_{11}, B_{22}^* upper Hessenberg matrices and A_{22}, B_{11}^* lower triangular, $B_{21}=0$, $A_{21}=\alpha e_n e_n^*$, is called *S-Hessenberg pencil.* Furthermore if $B_{11}=A_{22}=I$, then it is called *normalized S-Hessenberg pencil.* If A_{11}, B_{22}^*, A_{22}^*, B_{11} are upper triangular and B_{21}, $A_{21}=0$, then $A-\lambda B$ is called *S-triangular pencil.*

(ii) If A (or B) is invertible, then we say that $A^{-1}B$ ($B^{-1}A$) is an *S-Hessenberg matrix* if $A-\lambda B$ is an S-Hessenberg pencil and $A^{-1}B$ ($B^{-1}A$) is an *S-triangular matrix* if $A-\lambda B$ is an *S-triangular pencil.*

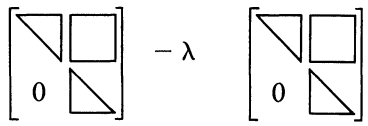
The structure of these pencils (matrices) can be easily described by the following diagrams:



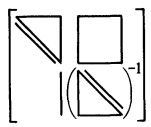
S-Hessenberg pencil,



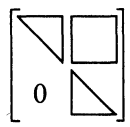
normalized S-Hessenberg pencil,



S-triangular pencil,



S-Hessenberg matrix,



S-triangular matrix.

Let

$$K:=\left[\begin{array}{c|c} I & 0 \\ \hline 0 & \begin{matrix} & 1 \\ \iddots & \\ 1 & \end{matrix}\end{array}\right]\in\mathbb{C}^{2n,2n},\qquad K=K^*=K^{-1}=J^TKJ. \tag{1.5}$$

Then if $A\in\mathbb{C}^{2n,2n}$ is S-triangular, then KAK is upper triangular.

Observe that if $A - \lambda B$ is a symplectic normalized S-Hessenberg pencil

$$\begin{bmatrix} F & 0 \\ H & I \end{bmatrix} - \lambda \begin{bmatrix} I & G \\ 0 & F^* \end{bmatrix},$$

then it follows that $GH = 0$.

2. Preliminaries. We now come back to our special pencil (0.6) arising in the discrete optimal control problem. Note that we assume from now on that $E = I$ and that rank $H = 1$; thus (0.6) is

$$\begin{bmatrix} F & 0 \\ H & I \end{bmatrix} - \lambda \begin{bmatrix} I & -G \\ 0 & F^* \end{bmatrix} =: \mathscr{A} - \lambda \mathscr{B}. \tag{2.1}$$

We have the following important observation.

PROPOSITION 2.2 (Cayley transformation). *Let $\mathscr{A} - \lambda \mathscr{B}$ be a symplectic pencil. Then for all $\lambda \in \mathbb{C} \backslash \sigma(\mathscr{A}, \mathscr{B})$,*

$$S = (\mathscr{A} - \lambda \mathscr{B})^{-1}(\bar{\lambda}\mathscr{A} - \mathscr{B}) \in S_{2n}. \tag{2.3}$$

Proof.

$$\begin{aligned} &(\bar{\lambda}\mathscr{A} - \mathscr{B})J(\bar{\lambda}\mathscr{A} - \mathscr{B})^* - (\mathscr{A} - \lambda\mathscr{B})J(\mathscr{A} - \lambda\mathscr{B})^* \\ &\quad = |\lambda|^2 \mathscr{A}J\mathscr{A}^* - \bar{\lambda}\mathscr{A}J\mathscr{B}^* - \lambda BJA^* + \mathscr{B}J\mathscr{B}^* \\ &\qquad - \mathscr{A}J\mathscr{A}^* + \lambda\mathscr{B}J\mathscr{A}^* + \bar{\lambda}AJB^* - |\lambda|^2 \mathscr{B}J\mathscr{B}^* = 0, \end{aligned} \tag{2.4}$$

since $\mathscr{A} - \lambda\mathscr{B}$ is a symplectic pencil. □

As in the usual QR- or QZ-algorithm, the algorithm that we will describe is based on a type of QR-decomposition. Here we consider the following factorization.

THEOREM 2.5. *Let $S \in S_{2n}$; then there exists a unique factorization $S = QT$, with $Q \in US_{2n}$ and $T \in S_{2n}$ is S-triangular, with positive diagonal elements.*

Proof. See Bunse-Gerstner [7] or Byers [4, p. 85] for the proof of this theorem. □

In the following we will denote this factorization as *QT-factorization.*

In order to motivate the algorithm that we will describe in a similar way as the QR- or QZ-algorithm (e.g., [13]), consider the following theorem.

THEOREM 2.6. *Let $\mathscr{A} - \lambda\mathscr{B}$ as in* (0.6) *be an S-Hessenberg pencil with $\mathscr{B}$ nonsingular. Let $\lambda_s \in \mathbb{C} \backslash \sigma(\mathscr{A}, \mathscr{B})$ and $\tilde{S} = (\mathscr{A} - \lambda_s\mathscr{B})^{-1}(\lambda_s\mathscr{A} - \mathscr{B})$. Let $Q \in US_{2n}$ such that $Q^*\tilde{S} = T \in S_{2n}$ is S-triangular. Then $Q^*\mathscr{B}^{-1}\mathscr{A}Q$ is an S-Hessenberg matrix.*

THEOREM 2.7 (Generalized symplectic Schur form). *Let $\mathscr{A} - \lambda\mathscr{B}$ be a regular symplectic pencil having no eigenvalues of modulus* 1. *Then there exists $Q \in U_{2n}$ and a matrix $Z \in US_{2n}$ such that*

$$Q\mathscr{A}Z = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} =: T, \tag{2.8}$$

$$Q\mathscr{B}Z = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} =: R, \tag{2.9}$$

and $T - \lambda R$ is an S-triangular pencil. $T_{11} - \lambda R_{11}$ has only eigenvalues λ with $|\lambda| < 1$ and $T_{22} - \lambda R_{22}$ has only eigenvalues λ with $|\lambda| > 1$ (including infinite eigenvalues). Furthermore letting $t_{jj}^{(1)}$, $t_{jj}^{(2)}$, $r_{jj}^{(1)}$, $r_{jj}^{(2)}$, $j = 1, \cdots, n$ be the diagonal elements of T_{11}, T_{22}, R_{11}, R_{22}, respectively, we have that

$$t_{jj}^{(1)}/r_{jj}^{(1)} = \overline{r_{jj}^{(2)}/t_{jj}^{(2)}}, \qquad j = 1, \cdots, n. \tag{2.10}$$

Proof. By the generalized Schur theorem of Stewart, e.g., [11, p. 269], there exist matrices $Q_1, Z_1 \in U_{2n}$ such that

$$Q_1(A-\lambda B)Z_1 = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ 0 & \Theta_{22} \end{bmatrix} - \lambda \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ 0 & \Psi_{22} \end{bmatrix}$$

with $\Theta_{11}, \Theta_{22}, \Psi_{11}, \Psi_{22}$ upper triangular, and we may assume that the eigenvalues are ordered such that $\Theta_{11} - \lambda\Psi_{11}$ contains all the eigenvalues λ with $|\lambda| < 1$. The eigenvalues of $\mathscr{A} - \lambda\mathscr{B}$, under the given assumptions, occur in pairs λ, $1/\lambda$ (Pappas, Laub, and Sandell [18]) and there are exactly n eigenvalues of modulus $|\lambda| < 1$. Now we play the "Laub-trick," e.g., instead of

$$Z_1 = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}$$

we consider

$$Z = \begin{bmatrix} Z_{11} & -Z_{21} \\ Z_{21} & Z_{11} \end{bmatrix}.$$

It has been shown by Pappas, Laub, and Sandell [18] that if $\left[\begin{smallmatrix} Z_{11} \\ Z_{21} \end{smallmatrix}\right]$ is the deflating subspace to $\mathscr{A} - \lambda\mathscr{B}$ corresponding to the eigenvalues of modulus less than one, then under the given assumptions, (I, A, B) stabilizable, $(I, A, 0)$ detectable, it follows that Z_{11}^{-1} exists and $-Z_{21}Z_{11}^{-1}$ is the Hermitian positive semidefinite solution of (0.9) (with $E = I$). Hence it follows that

$$\begin{bmatrix} Z_{11} & -Z_{21} \\ Z_{21} & Z_{11} \end{bmatrix}^* \begin{bmatrix} Z_{11} & -Z_{21} \\ Z_{21} & Z_{11} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \tag{2.11}$$

and

$$\begin{bmatrix} Z_{11} & -Z_{21} \\ Z_{21} & Z_{11} \end{bmatrix}^* \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \begin{bmatrix} Z_{11} & -Z_{21} \\ Z_{21} & Z_{11} \end{bmatrix} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}. \tag{2.12}$$

Here we use that Z_1 is unitary and $Z_{21}Z_{11}^{-1}$ is Hermitian. We thus have

$$(\mathscr{A} - \lambda\mathscr{B})Z = \begin{bmatrix} \Theta_{11} & \tilde{\Theta}_{12} \\ 0 & \tilde{\Theta}_{22} \end{bmatrix} - \lambda \begin{bmatrix} \Psi_{11} & \tilde{\Psi}_{12} \\ 0 & \tilde{\Psi}_{22} \end{bmatrix} \tag{2.13}$$

but $\tilde{\Theta}_{22}, \tilde{\Psi}_{22}$ are no longer triangular. Let $\tilde{Q}_{22}$ be unitary such that $\tilde{Q}_{22}\tilde{\Theta}_{22}$ is lower triangular; then we get that

$$\begin{bmatrix} I & 0 \\ 0 & \tilde{Q}_{22} \end{bmatrix} \left(\begin{bmatrix} \Theta_{11} & \tilde{\Theta}_{12} \\ 0 & \tilde{\Theta}_{22} \end{bmatrix} - \lambda \begin{bmatrix} \Psi_{11} & \tilde{\Psi}_{12} \\ 0 & \tilde{\Psi}_{22} \end{bmatrix} \right) = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} - \lambda \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} =: T - \lambda R.$$

Observe that we still have that $T - \lambda R$ is a symplectic pencil, i.e., $TJT^* = RJR^*$ and this implies $T_{11}T_{22}^* = R_{11}R_{22}^*$. $T_{11} - \lambda R_{11} = \Theta_{11} - \lambda\Psi_{21}$ has only eigenvalues λ with $|\lambda| < 1$; thus R_{11} is invertible. This implies that R_{22}^* is upper triangular. T_{22} is also invertible, since $T_{22} - R_{22}$ has only eigenvalues λ satisfying $|\lambda| > 1$; thus it follows that $R_{11}^{-1}T_{11} = R_{22}^*T_{22}^{-*}$.

This then implies (2.10). □

COROLLARY 2.14. *Under the assumptions of Theorem* 2.7, *there exist* $Q \in U_{2n}$, $Z \in US_{2n}$, $S_1 \in S_{2n}$, $Q_1 \in \mathbb{C}^{2n,2n}$ *such that* S_1 *is S-triangular,* Q_1 *is lower triangular and*

$$Q_1Q(\mathscr{A} - \lambda\mathscr{B})ZS_1 = \begin{bmatrix} T_{11} & T_{12} \\ 0 & I \end{bmatrix} - \lambda \begin{bmatrix} I & R_{12} \\ 0 & T_{11}^* \end{bmatrix}. \tag{2.15}$$

Proof. Following the proof of Theorem 2.7, we choose

$$S_1 = \begin{bmatrix} R_{11}^{-1} & 0 \\ 0 & R_{11}^* \end{bmatrix} \quad \text{and} \quad Q_1 = \begin{bmatrix} I & 0 \\ 0 & (T_{22}R_{11}^*)^{-1} \end{bmatrix}.$$

Both matrices exist by the proof of Theorem 2.7 and, since $ZS_1 \in S_{2n}$, (2.15) follows. □

COROLLARY 2.16. *Let $S \in S_{2n}$ having no eigenvalue of modulus one. Then there exists $Z \in US_{2n}$, such that*

$$Z^*SZ = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{11}^{-*} \end{bmatrix},$$

where T_{11} is upper triangular and has only eigenvalues λ with $|\lambda| < 1$.

Proof. Apply Theorem 2.7 to $S - \lambda I$ to obtain the proof. □

Remark 2.17. In general there do not exist matrices $Q \in U_{2n}$, $Z \in US_{2n}$ such that

$$Q(\mathscr{A} - \lambda\mathscr{B})Z = \begin{bmatrix} T_{11} & T_{12} \\ 0 & I \end{bmatrix} - \lambda \begin{bmatrix} I & R_{12} \\ 0 & T_{11}^* \end{bmatrix}.$$

It is not even possible to achieve a form

$$\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} - \lambda \begin{bmatrix} T_{22}^* & R_{12} \\ 0 & T_{11}^* \end{bmatrix}$$

in this way. To see this, consider the following example. Let

$$\mathscr{A} - \lambda\mathscr{B} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} - \lambda \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix},$$

let

$$Q_1 = \begin{bmatrix} c_1 & s_1 \\ -\bar{s}_1 & \bar{c}_1 \end{bmatrix} \in U_2,$$

$$Q_2 = \begin{bmatrix} c_2 & -s_2 \\ s_2 & c_2 \end{bmatrix} \in US_2,$$

i.e., $|c_1|^2 + |s_1|^2 = 1$, $|c_2|^2 + |s_2|^2 = 1$, $c_2\bar{s}_2 \in \mathbb{R}$. Suppose we have that

$$Q_1(\mathscr{A} - \lambda\mathscr{B})Q_2 = \begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} \end{bmatrix} - \lambda \begin{bmatrix} \bar{a}_{22} & b_{12} \\ 0 & \bar{a}_{11} \end{bmatrix}.$$

Then it follows that c_2, s_2, c_1, $s_1 \neq 0$ and

$$(c_1 + 2s_1)c_2 + s_1 s_2 = s_1\bar{s}_2 + \bar{c}_2(s_1 + c_1), \tag{2.18}$$

$$c_1 c_2 + s_2 s_1 - s_2 c_1 = c_1\bar{c}_2 - 2c_1\bar{s}_2 + s_1\bar{s}_2, \tag{2.19}$$

$$(-\bar{s}_1 + 2\bar{c}_1)c_2 + \bar{c}_1 s_2 = 0, \tag{2.20}$$

$$-\bar{s}_1 c_2 + (\bar{s}_1 + \bar{c}_1)s_2 = 0, \tag{2.21}$$

(2.20) and (2.21) imply

$$2\bar{c}_1 c_2 + \bar{s}_1 s_2 = 0. \tag{2.22}$$

Multiplying (2.22) by $\bar{s}_2 c_1$ and using $c_2\bar{s}_2 \in \mathbb{R}$, we obtain that $c_1\bar{s}_1 \in \mathbb{R}$, too. Multiplying (2.18) by $\bar{c}_1$ and ordering yields

$$|c_1|^2(c_2 - \bar{c}_2) + \bar{c}_1 s_1(s_2 - \bar{s}_2) + \bar{c}_1 s_1(2c_2 - \bar{c}_2) = 0. \tag{2.23}$$

The first two terms are purely imaginary; hence $2c_2 - \bar{c}_2 = i\alpha$, $\alpha \in \mathbb{R}$, but this implies that $c_2 \in i\mathbb{R}$ and hence also $s_2 \in i\mathbb{R}$. Combining (2.19), (2.22) we obtain

$$c_1 c_2 + s_2 s_1 - s_2 c_1 = -c_1 \bar{c}_2 - 2c_1 \bar{s}_2 \tag{2.24}$$

and this implies

$$s_2 s_1 = c_1(-2\bar{s}_2 + s_2). \tag{2.25}$$

Since s_2, $c_2 \in i\mathbb{R}$ it follows that

$$s_2(3c_1 - s_1) = 0. \tag{2.26}$$

If $s_2 = 0$, then $c_2 = \pm i$, $c_1 = 0$, $s_1 = \pm i$ and by (2.18) $\pm 1 = \pm 2$, which is impossible. So if $3c_1 = s_1$, then by (2.20) we obtain

$$-\bar{c}_1 c_2 + \bar{c}_1 s_2 = 0. \tag{2.27}$$

Thus, $\bar{c}_1 = 0$ or $c_2 = s_2$. If $\bar{c}_1 = 0$, then $\bar{s}_1 = 0$, which is impossible since $|s_1|^2 + |c_1|^2 = 1$. If $c_2 = s_2$, then by (2.21) we have $-\bar{s}_1 c_2 + \bar{s}_1 c_2 + \bar{c}_1 c_2 = 0$, which again leads to a contradiction. □

In the real case, we have to replace triangular by quasitriangular forms in these Schur forms if we want a real Schur form.

THEOREM 2.28 (Generalized symplectic real Schur form). *Let $A - \lambda B$ be a real regular symplectic pencil having no eigenvalues of modulus 1. Then there exists $Q \in U_{2n}$ and $Z \in US_{2n}$, Q, Z real, such that*

$$Q\mathscr{A}Z = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} =: T, \tag{2.29}$$

$$Q\mathscr{B}Z = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} =: R \tag{2.30}$$

and $T - \lambda R$ is a real quasitriangular pencil, i.e., $T_{11} - \lambda R_{11}$, $T_{22}^* - \lambda R_{22}^*$ are quasitriangular matrices (i.e., block triangular matrices with 1×1 or 2×2 diagonal blocks). $T_{11} - \lambda R_{11}$ has only eigenvalues λ with $|\lambda| < 1$ and $T_{22} - \lambda R_{22}$ has only eigenvalues λ with $|\lambda| > 1$ (including infinite eigenvalues) and further if $\tau_{jj}^{(1)}$, $\tau_{jj}^{(2)}$, $\rho_{jj}^{(1)}$, $\rho_{jj}^{(2)}$, $j = 1, \cdots, k$ are the 1×1 or 2×2 diagonal blocks of $T_{ii} - \lambda R_{ii} i = 1, 2$, respectively, then

$$\tau_{jj}^{(1)} \cdot (\rho_{jj}^{(1)})^{-1} = (\rho_{jj}^{(2)} \cdot (\tau_{jj}^{(2)})^{-1})^*, \qquad j = 1, \cdots, k. \tag{2.31}$$

Proof. The proof is analogous to that of Theorem 2.7.

Clearly, results analogous to Corollaries 2.14 and 2.16 for the corresponding real Schur form hold too.

In order to perform a double shift algorithm in the case of real $\mathscr{A}$, $\mathscr{B}$ and complex shifts, we need the following result.

PROPOSITION 2.32. *Let $A - \lambda B$ be a real symplectic pencil with B invertible. Let $s \in \mathbb{C} \backslash \sigma(A, B)$ and let $Q_1 T_1$ be the QT-factorization of $(A - sB)^{-1}(sA - B)$. Let $A_1 - \lambda B_1 := (A - \lambda B)Q_1$. Let $Q_2 T_2$ be the QT-factorization of $(A_1 - \bar{s}B_1)^{-1}(\bar{s}A_1 - B_1)$, then*

$$QT := Q_1 Q_2 T_2 T_1 = (A - \bar{s}B)^{-1} B (A - sB)^{-1} (\bar{s}A - B) B^{-1} (sA - B) \tag{2.33}$$

is a real QT-factorization.

Proof.

$$\begin{aligned} Q_1 Q_2 T_2 T_1 &= Q_1(A_1 - \bar{s}B_1)^{-1}(\bar{s}A_1 - B_1)T_1 = (A - \bar{s}B)^{-1}(\bar{s}A - B)Q_1 T_1 \\ &= (A - \bar{s}B)^{-1}(\bar{s}A - B)(A - sB)^{-1}(sA - B) \\ &= (A - \bar{s}B)^{-1}(\bar{s}AB^{-1} - I)(AB^{-1} - sI)^{-1}(sA - B) \\ &= (A - \bar{s}B)^{-1}B(A - sB)^{-1}(\bar{s}A - B)(sA - B) \\ &= [(B^{-1}A - sI)(B^{-1}A - \bar{s}I)]^{-1}(\bar{s}B^{-1}A - I)(sB^{-1}A - I) \end{aligned}$$

which is real since the two indicated factors are real. Consequently the uniqueness of the QT-factorization yields that Q_1, Q_2, $T_2 T_1$ are real factors. □

We immediately then have the following.

COROLLARY 2.34. *Let M be a real symplectic matrix. Let $s \in \mathbb{C}$ and let $Q_1 T_1$ be the QT-factorization of*

$$(M - sI)^{-1}(sM - I).$$

Let $M_1 = Q_1^* M_1 Q_1$ and let $Q_2 T_2$ be the QT-factorization of $(M - \bar{s}I)^{-1}(\bar{s}M - I)$; then

$$QT = Q_1 Q_2 T_2 T_1 = (M - \bar{s}I)^{-1}(M - sI)^{-1}(\bar{s}M - I)(sM - I) \tag{2.35}$$

is a real QT-factorization.

Another important tool in the study of symplectic pencils/matrices is the following.

PROPOSITION 2.36. *Let*

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \in S_{2n},$$

and suppose that S_{22}^{-1} exists. Then S can be factored into the following product of three symplectic factors:

$$S = \begin{bmatrix} I & S_{12}S_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} S_{22}^{-*} & 0 \\ 0 & S_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ S_{22}^{-1}S_{21} & I \end{bmatrix}. \tag{2.37}$$

Proof. We only have to show that

$$S_{22}^{-*} + S_{12}S_{22}^{-1}S_{21} = S_{11}.$$

But S is symplectic; thus

$$S_{21}S_{22}^{*} = S_{22}S_{21}^{*}, \qquad S_{11}S_{22}^{*} - S_{12}S_{21}^{*} = I,$$

which implies

$$S_{22}^{-*} + S_{12}S_{22}^{-1}S_{21} = S_{22}^{-*} + S_{12}S_{21}^{*}S_{22}^{-*} = S_{22}^{-*} + S_{11}S_{22}^{*}S_{22}^{-*} - S_{22}^{-*} = S_{11}. \qquad \square$$

Note, that if S_{11} is invertible, then we obtain the analogous factorization

$$S = \begin{bmatrix} I & 0 \\ S_{21}S_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} S_{11} & 0 \\ 0 & S_{11}^{-*} \end{bmatrix} \begin{bmatrix} I & S_{11}^{-1}S_{12} \\ 0 & I \end{bmatrix}. \tag{2.38}$$

3. The algorithms. Given a symplectic pencil

$$\begin{bmatrix} F & 0 \\ H & I \end{bmatrix} - \lambda \begin{bmatrix} \mathrm{I} & -G \\ 0 & F^* \end{bmatrix} =: \mathscr{A} - \lambda\mathscr{B} =: [a_{ij}] - \lambda[b_{ij}] \tag{3.1}$$

or a symplectic matrix

$$\mathcal{M} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = [m_{ij}] \tag{3.2}$$

where $F, G, H \in \mathbb{C}^{n,n}$, $H = H^*$, $G = G^*$, rank $(H) = 1$, rank $(M_{21}) = 1$, and H, G positive semidefinite. (See the Introduction.) We now describe algorithms for the computation of the deflating (invariant) subspaces of $\mathcal{A} - \lambda\mathcal{B}(\mathcal{M})$ corresponding to the eigenvalues λ, with $|\lambda| < 1$. We begin with the description of the preliminary reduction step to S-Hessenberg form.

ALGORITHM 3.3 (Reduction to S-Hessenberg form).

Step 1. Simplification of $H(M_{21})$. Let $u \in \mathbb{C}^n$ such that $H = uu^*(u^* = e_n^* M_{21})$. Let $U \in U_n$ such that

$$U^* u = \alpha e_n.$$

Let $Q_0 = \left[\begin{smallmatrix} U & 0 \\ 0 & U \end{smallmatrix}\right]$, then

$$\begin{aligned} Q_0^*(\mathcal{A} - \lambda\mathcal{B})Q_0 &= \begin{bmatrix} U^*FU & 0 \\ |\alpha|^2 e_n e_n^* & I \end{bmatrix} - \lambda \begin{bmatrix} I & -U^*GU \\ 0 & U^*F^*U \end{bmatrix} \\ &=: \begin{bmatrix} F_0 & 0 \\ H_0 & I \end{bmatrix} - \lambda \begin{bmatrix} I & -G_0 \\ 0 & F_0^* \end{bmatrix} =: A_0 - \lambda B_0. \end{aligned} \tag{3.4}$$

$$Q_0^* \mathcal{M} Q_0 =: \begin{bmatrix} M_{11}^{(0)} & M_{12}^{(0)} \\ M_{21}^{(0)} & M_{22}^{(0)} \end{bmatrix} =: M_0 \quad \text{and} \quad M_{21}^{(0)} \triangleq \begin{bmatrix} & & \\ & 0 & \\ & & \end{bmatrix}. \tag{3.5}$$

Set $Q := Q_0$.

U can be obtained in the usual way using a Householder transformation, e.g., [13].

Step 2. Reduction of $F_0(M_{11}^{(0)})$. Let $V \in U_n$ such that $F_1 := V^* F_0 V (V^* M_{11}^{(0)} V)$ is upper Hessenberg and

$$H_1 := V^* H_0 V = H_0 \left(M_{21}^{(1)} := M_{21}^{(0)} \triangleq \begin{bmatrix} & & \\ & 0 & \\ & & \end{bmatrix} \right).$$

This is achieved by the usual transformation to Hessenberg form, e.g., [13], here with the last column of V fixed to be e_n. Set $G_1 = V^* G_0 V$ and $Q_1 =: \left[\begin{smallmatrix} V & 0 \\ 0 & V \end{smallmatrix}\right]$. Then

$$Q_1^*(A_0 - \lambda B_0)Q_1 = \begin{bmatrix} F_1 & 0 \\ H_1 & I \end{bmatrix} - \lambda \begin{bmatrix} I & -G_1 \\ 0 & F_1^* \end{bmatrix} =: A_1 - \lambda B_1 \tag{3.6}$$

is an S-Hessenberg pencil and

$$Q_1^* M Q_1 = \begin{bmatrix} M_{11}^{(1)} & M_{12}^{(1)} \\ M_{21}^{(1)} & M_{22}^{(1)} \end{bmatrix} =: M_1 \tag{3.7}$$

is an S-Hessenberg matrix.

Set $Q := Q \cdot Q_1$.

Observe that with these initial reductions we have produced the following structures:

(3.8) $$A_1 - \lambda B_1 \triangleq \begin{bmatrix} \ddots & 0 \\ * & \ddots \end{bmatrix} - \lambda \begin{bmatrix} \ddots & \square \\ 0 & \ddots \end{bmatrix},$$

(3.9) $$M_1 \triangleq \begin{bmatrix} \ddots & \square \\ \square & \ddots \end{bmatrix}^{-1} = \begin{bmatrix} \ddots & \square \\ \square & \ddots \end{bmatrix}$$

where in the following we denote by "$\triangleq$" that the matrix is of that given form.

By Theorem 2.6 we have that the form of M_1 stays invariant if we perform the following QR-type iteration which is similar to the Hamiltonian QR-iteration of Byers [5]. It is known that for general symplectic matrices, i.e., if they do not arise from a single input optimal control problem as above, the simplest Hessenberg type form that we can obtain with unitary symplectic transformation is

$$\begin{bmatrix} \ddots & \square \\ \ddots & \square \end{bmatrix},$$

which is not an invariant form, under the symplectic QR-type iteration, e.g., Bunse-Gerstner [7]. The iterative part can be carried out by the following algorithms.

THE *DSQR* ALGORITHM 3.10 (*D*iscrete *S*ymplectic *QR*-algorithm). (Single or double shift step). Given M_i an S-Hessenberg matrix as in (3.7), (3.9), and $Q \in US_{2n}$.

FOR $i = 1, 2, 3, \cdots$
Choose a shift $\lambda_i \in \mathbb{C}$
For a single shift step let

(3.11) $S_i = (M_i - \lambda_i I)^{-1}(\bar{\lambda}_i M_i - I)$
and for a double shift step let

(3.12) $S_i = (M_i - \bar{\lambda}_i I)^{-1}(M_i - \lambda_i I)^{-1}(\bar{\lambda}_i M_i - I)(\lambda_i M_i - I)$.
Compute a QT-factorization of S_i, $S_i = U_i T_i$ as in Theorem 2.5 and set

(3.13) $M_{i+1} = U_i^* M_i U_i$
Set

(3.14) $Q := QU_i$.
END

It is also possible to perform a QZ-type algorithm, but as we have seen in Remark 2.17, it is not possible to produce the Schur form using unitary transformations from the left. This is also the case in this algorithm.

THE *DSSZ* ALGORITHM 3.15 (*D*iscrete *S*ymplectic *SZ*-algorithm). (Single or double shift version). Given $A_1 - \lambda B_1$ an S-Hessenberg pencil as in (3.6), (3.8), and $Q \in US_{2n}$.

FOR $i = 1, 2, 3, \cdots$
Choose a shift $\lambda_i \in \mathbb{C}$.
⊙ For a single shift step let

(3.16) $S_i = (A_i - \lambda B_i)^{-1}(\bar{\lambda}_i A_i - B_i)$
and for a double shift step let
(3.17) $S_i = (A_i - \lambda B_i)^{-1} B (A_i - \bar{\lambda} B_i)^{-1}(\bar{\lambda}_i A_i - B_i) B^{-1}(\lambda_i A_i - B_i)$.
Compute a QT-factorization of S_i, $S_i = U_i T_i$ as in Theorem 2.5, and let $Z_i \in \mathbb{C}^{n,n}$ nonsingular, such that
(3.18) $A_{i+1} - B_{i+1} = Z_i A_i U_i - \lambda Z_i B_i U_i$
is again a normalized S-Hessenberg pencil. If such a Z_i does not exist, choose a different shift λ_i and GOTO ⊙. Set
(3.19) $Q = Q \cdot U_i$
END

In general we may not assume that F (and thus B) is invertible, so the matrix B^{-1} may not exist. But it is possible to deflate zero and infinite eigenvalues of $A - \lambda B$ so that during the iteration, we may assume that the iterates A_i, B_i are nonsingular and also that in general we are able to form the matrix $M = B^{-1}A$.

Let

$$A - \lambda B = \begin{bmatrix} F & 0 \\ H & I \end{bmatrix} - \lambda \begin{bmatrix} I & -G \\ 0 & F^* \end{bmatrix},$$

with F upper Hessenberg, $H = \alpha e_n e_n^*$, $G = G^*$, G, H positive semidefinite, and F unreduced, i.e., all subdiagonal elements of F are nonzero. (This can be achieved by the deflation procedure described later.) Then we consider the following.

ALGORITHM 3.20 Deflation of eigenvalues 0, ∞. Let

$$T_1 = \begin{bmatrix} I & 0 \\ H & I \end{bmatrix}, \quad T_2 = \begin{bmatrix} Z^* & 0 \\ 0 & Z \end{bmatrix}, \quad T_3 = \begin{bmatrix} I & 0 \\ 0 & Z^* \end{bmatrix},$$

where $Z \in U_n$, such that $FZ = R$ is upper triangular. Then,

$$\begin{aligned} T_3(A - \lambda B) T_1^{-1} T_2^* &= T_3 \left(\begin{bmatrix} F & 0 \\ 0 & I \end{bmatrix} - \lambda \begin{bmatrix} I + GH & -G \\ -F^*H & F^* \end{bmatrix} \right) \begin{bmatrix} Z & 0 \\ 0 & Z \end{bmatrix} \\ &= \begin{bmatrix} R & 0 \\ 0 & I \end{bmatrix} - \lambda \begin{bmatrix} (I + GH)Z & -GZ \\ -R^*HZ & Z^*F^*Z \end{bmatrix}. \end{aligned} \tag{3.21}$$

Now H, G are positive semidefinite; thus $(I + GH)^{-1}$ exists. Let

$$T_4 = \begin{bmatrix} Z(I + GH)^{-1} & 0 \\ 0 & I \end{bmatrix}, \quad T_5 = \begin{bmatrix} I & 0 \\ R^*HZ & I \end{bmatrix}.$$

Then

$$\begin{aligned} A - \lambda B &:= T_5 T_4 T_3 (A - \lambda B) T_1^{-1} T_2^* \\ &= \begin{bmatrix} Z^*(I + GH)^{-1} R & 0 \\ R^*H(I + GH)^{-1} R & I \end{bmatrix} - \lambda \begin{bmatrix} I & -Z^*(I + GH)^{-1} GZ \\ 0 & R^*Z - R^*H(I + GH)^{-1} GZ \end{bmatrix}. \end{aligned} \tag{3.22}$$

If F is singular then R has first column identically zero and then

(3.23) $$\tilde{A}-\lambda\tilde{B}\triangleq \left[\begin{array}{c|c} \begin{matrix}0\\ \vdots\\ 0\end{matrix} \quad \square & 0 \\ \hline 0 \quad \cdots \quad 0 & 1 \\ 0 \quad * \cdots * & \ddots \\ & 1 \end{array}\right] - \lambda \left[\begin{array}{c|c} \begin{matrix}1 & \\ \ddots & \\ & 1\end{matrix} \quad \square & \\ \hline 0 & \begin{matrix}0 \cdots \cdots 0\\ *\,*\,*\cdots 0\\ \vdots \quad \ddots \quad *\\ * \cdots \cdots *\end{matrix} \end{array}\right].$$

Delete rows and columns 1, $n+1$ and obtain a symplectic pencil of dimension 2 less

(3.24) $$\hat{A}-\lambda\hat{B}:=\begin{bmatrix}\hat{F}_1 & 0\\ \hat{H} & I\end{bmatrix}-\lambda\begin{bmatrix}I & -\hat{G}\\ 0 & F_2^*\end{bmatrix}\triangleq \left[\begin{array}{c|c} \ddots & 0 \\ \hline & \ddots \end{array}\right] - \lambda \left[\begin{array}{c|c} \ddots & \square \\ \hline 0 & \ddots \end{array}\right].$$

Since it is symplectic it follows that $\hat{H} = \beta e_{n-1}e_{n-1}^*$ and $\hat{F}_2 = \hat{F}_1$ and now $\hat{B}$ is non-singular.

Remark 3.25. Observe that the matrices $T_1, \cdots, T_5$ can also be obtained if instead of a symplectic pencil a symplectic S-Hessenberg matrix M is given. Using Proposition 2.36, M can be written as

(3.26) $$M=[m_{ij}]=\begin{bmatrix}M_{11} & M_{12}\\ M_{21} & M_{22}\end{bmatrix}=\begin{bmatrix}F-GF^{-*}H & GF^{-*}\\ F^{-*}H & F^{-*}\end{bmatrix},$$

where F, G, H are as above.

It is immediate to obtain $H = \alpha e_n e_n^*$ by comparing the last columns of M_{21}, M_{22}. Then

(3.27) $$\alpha = m_{2n,n}/m_{2n,2n},$$

(3.28) $$F = M_{11} - M_{12}\alpha e_n e_n^*,$$

and

(3.29) $$GH = M_{12}F^*H.$$

4. Detailed description of the algorithm. In this section we describe in detail the QT-factorizations used in the $DSQR$-algorithm, the implicit computation of the next iterate, the choice of shifts, deflation, the ordering of eigenvalues, and the determination of the deflating or invariant subspace.

The QT-factorization of $(M - sI)^{-1}(\bar{s}M - I)$. We now describe the procedure to obtain the QT-factorization of $S = (A - sB)^{-1}(\bar{s}A - B) = (M - sI)^{-1}(\bar{s}M - I)$, where A, B, S, M are iterates of the $DSQR$-algorithms and s is a shift parameter.

Let

(4.1) $$A=\begin{bmatrix}F & 0\\ H & I\end{bmatrix},\quad B=\begin{bmatrix}I & -G\\ 0 & F^*\end{bmatrix},\quad C:=\begin{bmatrix}I & 0\\ 0 & F^*\end{bmatrix}.$$

Assume further that F is unreduced, i.e., all subdiagonal elements are nonzero, and that $H \neq 0$. Then,

$$M = B^{-1}A = \begin{bmatrix} I & G \\ 0 & I \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & F^{-*} \end{bmatrix}\begin{bmatrix} F & 0 \\ H & I \end{bmatrix} = C^{-1}\begin{bmatrix} F+GF^{-*}H & GF^{-*} \\ H & I \end{bmatrix} =: C^{-1}\tilde{M}. \tag{4.2}$$

Clearly

$$S = (A - sB)^{-1}(\bar{s}A - B) = (M - sI)^{-1}(\bar{s}M - I) = (\tilde{M} - sC)^{-1}(s\tilde{M} - C),$$

$$\tilde{M} - SC = \begin{bmatrix} F - sI + GF^{-*}H & GF^{-*} \\ H & I - sF^* \end{bmatrix} =: [\tilde{m}_{i,j}] - s[c_{ij}]. \tag{4.3}$$

Let $Z_1 R_1$ be a QR-factorization of $\bar{s}F - I$ (e.g., Golub and van Loan [13] or Bunse and Bunse-Gerstner [11]).

Observe that

$$GF^{-*}H := ue_n^* \triangleq \begin{bmatrix} & & \\ & 0 & \\ & & \end{bmatrix}$$

for some $u \in \mathbb{C}^n$. Thus, $Z_1^*(\bar{s}(F + GF^{-*}H) - I) = R_2$ is also upper triangular, and also we have that Z_1 is upper Hessenberg. Then

$$W_1 := [w_{ij}^{(1)}] := \begin{bmatrix} W_{11}^{(1)} & W_{12}^{(1)} \\ W_{21}^{(1)} & W_{21}^{(1)} \end{bmatrix} = \begin{bmatrix} Z_1^* & 0 \\ 0 & I \end{bmatrix}(\tilde{M} - sC)\begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix} \triangleq \begin{bmatrix} & \\ & \\ ** & \end{bmatrix} \tag{4.4}$$

and

$$Y_1 := [y_{i,j}^{(1)}] := \begin{bmatrix} Z_1^* & 0 \\ 0 & I \end{bmatrix}(\bar{s}\tilde{M} - C) \triangleq \begin{bmatrix} & \\ & \\ * & \end{bmatrix} \tag{4.5}$$

and

$$\begin{bmatrix} Z_1^* & 0 \\ 0 & Z_1^* \end{bmatrix} S = W_1^{-1} Y_1.$$

Observe that the structures of W_1, Y_1 follow, since

$$W_{11}^{(1)} = Z_1^*(F - sI - GF^{-*}H)Z_1$$

$$= \frac{1}{\bar{s}}(\bar{s}Z_1^*(F - GF^{-*}H)Z_1 - I) + \left(\frac{1}{\bar{s}} - s\right)I = \frac{1}{\bar{s}}R_2 Z_1 + \left(\frac{1}{\bar{s}} - s\right)I$$

is upper Hessenberg, $W_{21}^{(1)} = HZ_1$ and $W_{22}^{(1)} = (I - sF^*)Z_1 = (Z_1^*(I - \bar{s}F))^* = -R_1^*$ is lower triangular. Now let $J_1 = J(n, c_1, s_1)$ be Jacobi symplectic, such that

$$\begin{bmatrix} c_1 & s_1 \\ -s_1 & c_1 \end{bmatrix}\begin{bmatrix} y_{n,n}^{(1)} \\ y_{2n,n}^{(1)} \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix}. \tag{4.6}$$

Then

$$Y_2 := [y^{(2)}_{i,j}] = J_1 Y_1 \triangleq \begin{bmatrix} \blacksquare & \square \\ 0 & \blacksquare \end{bmatrix} \tag{4.7}$$

and

$$W_2 := [w^{(2)}_{i,j}] = J_1 W_1 \triangleq \begin{bmatrix} \blacksquare & \square \\ * & \blacksquare \end{bmatrix}. \tag{4.8}$$

The element $w^{(2)}_{2n,n-1}$ vanishes automatically, since with $R_1 := [r_{ij}]$, $Z_1 = [z_{ij}]$ we have

$$\begin{aligned} &Y^{(1)}_{n,n} = r_{n,n}, y^{(1)}_{2n,n} = \bar{s}\tilde{m}_{2n,n}, w^{(1)}_{n,n-1} = \frac{1}{\bar{s}} r_{n,n} z_{n,n-1}, \\ &W^{(1)}_{2n,n-1} = \tilde{m}_{2n,n} z_{n,n-1}. \end{aligned} \tag{4.9}$$

Thus, the choice of c_1, s_1 as in (4.6) simultaneously eliminates $y^{(1)}_{2n,n}$, $w^{(1)}_{2n,n-1}$.

Now let $J_2 = J(n, c_2, s_2) \in US_{2n}$ such that

$$[w^{(2)}_{2n,n} w^{(2)}_{2n,n}] \begin{bmatrix} c_2 & s_2 \\ -s_2 & c_2 \end{bmatrix} = [0 \quad *].$$

Then

$$W_3 =: \begin{bmatrix} W^{(3)}_{11} & W^{(3)}_{12} \\ W^{(3)}_{21} & W^{(3)}_{22} \end{bmatrix} := W_2 J_2 \triangleq \begin{bmatrix} \blacksquare & \square \\ 0 & \blacksquare \end{bmatrix}.$$

Choose $Z_2 \in U_n$ such that $W^{(3)}_{11} Z_2 = R_3$ is upper triangular and set

$$W_4 := W_3 \begin{bmatrix} Z_2 & 0 \\ 0 & Z_2 \end{bmatrix}.$$

Then

$$W_4 \triangleq \begin{bmatrix} \blacksquare & \square \\ 0 & \square \end{bmatrix}.$$

Let

$$U := \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix} J_2 \begin{bmatrix} Z_2 & 0 \\ 0 & Z_2 \end{bmatrix} \quad \text{and} \quad Z := J_1 \begin{bmatrix} Z_1^* & 0 \\ 0 & I \end{bmatrix}.$$

$$\begin{aligned} U^*(A - sB)^{-1}(\bar{s}A - B) &= U^*(C^{-1}A - sC^{-1}B)^{-1} Z^* Z(\bar{s}C^{-1}A - C^{-1}B) \\ &= [Z(C^{-1}A - sC^{-1}B)U]^{-1}[Z(\bar{s}C^{-1}A - C^{-1}B)] =: T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \end{aligned}$$

has the structure

$$\begin{bmatrix} \blacksquare & \square \\ 0 & \square \end{bmatrix}^{-1} \begin{bmatrix} \blacksquare & \square \\ 0 & \square \end{bmatrix} = \begin{bmatrix} \blacksquare & \square \\ 0 & \square \end{bmatrix}.$$

But since $U \in \mathcal{U}S_{2n}$, it follows that $T \in S_{2n}$ and thus $T_{22} = T_{11}^{-*}$, i.e., we have produced the unique QT-factorization of $(A - sB)^{-1}(\bar{s}A - B)$.

In a similar way, we now obtain the QT-factorization in the double shift procedure.

The QT-factorization of $S = (M - \bar{s}I)^{-1}(M - sI)^{-1}(\bar{s}M - I)(sM - I)$. Let

$$M = \begin{bmatrix} I & -G \\ 0 & F^* \end{bmatrix}^{-1} \begin{bmatrix} F & 0 \\ H & I \end{bmatrix} = B^{-1}A.$$

Then

$$\begin{aligned}
(A - \bar{s}B)B^{-1}(A - sB) &= B(M - \bar{s}I)(M - sI) \\
&= \begin{bmatrix} F - \bar{s}I & sG \\ H & I - \bar{s}F^* \end{bmatrix} \begin{bmatrix} I & GF^{-*} \\ 0 & F^{-*} \end{bmatrix} \begin{bmatrix} F - sI & sG \\ H & I - sF^* \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ 0 & F^{-*} \end{bmatrix} \begin{bmatrix} F - \bar{s}I & sGF^{-*} \\ F^*H & I - \bar{s}F^* \end{bmatrix} \begin{bmatrix} F - sI + GF^{-*}H & GF^{-*} \\ H & I - sF^* \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ 0 & F^{-*} \end{bmatrix} \begin{bmatrix} (F - \bar{s}I)(F - sI) + FGF^{-*}H & FGF^{-*} - |s|^2 G \\ F^*HF - \bar{s}F^*H + F^*HGF^{-*}H + H - sF^*H & F^*HGF^{-*} + (I - \bar{s}F^*)(I - sF^*) \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ 0 & F^{-*} \end{bmatrix} L_1,
\end{aligned} \tag{4.10}$$

and

$$\begin{aligned}
(\bar{s}A - B)^{-1}(sA - B) &= \begin{bmatrix} I & 0 \\ 0 & F^{-*} \end{bmatrix} \begin{bmatrix} (\bar{s}F - I)(sF - I) + |s|^2 FGF^{-*}H & |s|^2 FGF^{-*} - G \\ |s|^2(F^*HF + F^*HGF^{-*}H + H) - 2\,\mathrm{Re}\,(s)F^*H & |s|^2 F^*HGF^{-*} + (sI - F^*)(sI - F^*) \end{bmatrix} \\
&:= \begin{bmatrix} I & 0 \\ 0 & F^{-*} \end{bmatrix} L_2.
\end{aligned} \tag{4.11}$$

Observe that

$$L_1 \triangleq \begin{bmatrix} \blacktriangle & \square \\ ** & \square \end{bmatrix} \triangleq L_2. \tag{4.12}$$

Let $(\bar{s}F - I)(sF - I) = Z_1 R_1$ be a QR-factorization, then

$$\begin{bmatrix} Z_1^* & 0 \\ 0 & I \end{bmatrix} L_2 = Y_1 = [y_{i,j}^{(1)}] = \begin{bmatrix} \blacktriangle & \square \\ ** & \square \end{bmatrix}. \tag{4.13}$$

Let

$$J_1 = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & c_1 & s_1 \\ & & & -s_1 & c_1 \end{bmatrix} \in U_{2n}$$

such that

$$\begin{bmatrix} c_1 & s_1 \\ -s_1 & c_1 \end{bmatrix}\begin{bmatrix} y^{(1)}_{2n-1,n-1} \\ y^{(1)}_{2n,n-1} \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix} \tag{4.14}$$

and let

$$Y_2 = [y^{(2)}_{i,j}] = J_1 Y_1.$$

Let $J_2 = J(n-1, c_2, s_2) \in US_{2n}$ such that

$$\begin{bmatrix} c_2 & s_2 \\ -s_2 & c_s \end{bmatrix}\begin{bmatrix} y^{(2)}_{n-1,n-1} \\ y^{(2)}_{2n-1,n-1} \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix} \tag{4.15}$$

and let

$$Y_3 = [y^{(3)}_{i,j}] = J_2 Y_2. \tag{4.16}$$

Let

$$J_3 = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & c_3 & s_3 \\ & & & -s_3 & c_3 \end{bmatrix} \in U_{2n}$$

such that

$$\begin{bmatrix} c_3 & s_3 \\ -s_3 & c_3 \end{bmatrix}\begin{bmatrix} y^{(3)}_{2n-1,n} \\ y^{(3)}_{2n,n} \end{bmatrix} = \begin{bmatrix} 0 \\ * \end{bmatrix} \tag{4.17}$$

and let

$$Y_4 = J_3 Y_3 = [y^{(4)}_{i,j}]. \tag{4.18}$$

Let $J_4 = J(n, c_4, s_4) \in US_{2n}$ such that

$$\begin{bmatrix} c_4 & s_4 \\ -s_4 & c_4 \end{bmatrix}\begin{bmatrix} y^{(4)}_{n,n} \\ y^{(4)}_{2n,n} \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix}. \tag{4.19}$$

Then

$$Y_5 = J_4 Y_4 \triangleq \begin{bmatrix} \blacktriangle & \square \\ 0 & \diagup\!\!\diagup \end{bmatrix}. \tag{4.20}$$

Let

$$W_1 := J_4 J_3 J_2 J_1 \begin{bmatrix} Z_1 & 0 \\ 0 & I \end{bmatrix} L_1 \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix} =: \begin{bmatrix} W^{(1)}_{21} & W^{(1)}_{12} \\ W^{(1)}_{21} & W^{(1)}_{22} \end{bmatrix} =: [w^{(1)}_{ij}]. \tag{4.21}$$

Observe that

$$[F^*HGF^{-*} + (I - sF^*)(I - \bar{s}F^*)]Z_1 \triangleq \begin{bmatrix} 0 \\ = \end{bmatrix} + \begin{bmatrix} \blacktriangle \\ \end{bmatrix} = \begin{bmatrix} \blacktriangle \\ * \end{bmatrix},$$

and thus

$$w_{22}^{(1)} = \begin{bmatrix} & \\ & \ddots \\ * & \end{bmatrix}.$$

Furthermore,

$$F^*HF - 2\,\mathrm{Re}\,(s)F^*H + F^*HGF^{-*}H \triangleq \begin{bmatrix} 0 & \\ & \begin{matrix} ** \\ ** \end{matrix} \end{bmatrix}.$$

Thus

$$W_1 \triangleq \begin{bmatrix} & \\ & \\ \begin{matrix} **** \\ **** \end{matrix} & * \end{bmatrix}. \tag{4.22}$$

Now $W_1^{-1}Y_5$ is symplectic. Thus, it follows that

$$W_{21}^{(1)}W_{22}^{(1)*} = W_{22}^{(1)}W_{21}^{(1)*} \tag{4.23}$$

and since L_1 was assumed invertible, the diagonal elements $w_{ii}^{(1)}$, $i = n+1, 2n-2$ are nonzero. Thus, we obtain that

$$W_1 \triangleq \begin{bmatrix} & \\ & \\ \begin{matrix} ** \\ ** \end{matrix} & * \end{bmatrix}. \tag{4.24}$$

Let

$$J_5 = \left[\begin{array}{ccc|ccc} 1 & & & & & \\ & \ddots & & & 0 & \\ & & 1 & & & \\ & & & c_5 & s_5 & \\ & & & -s_5 & c_5 & \\ \hline & & & & & 1 \\ & 0 & & & & \ddots \\ & & & & & 1 \\ & & & & & & c_5 & s_5 \\ & & & & & & -s_5 & c_5 \end{array}\right] \in US_{2n}$$

such that

$$[w_{2n-1,n-1}^{(1)}, w_{2n-1,n}^{(1)}]\begin{bmatrix} c_5 & s_5 \\ s_5 & c_5 \end{bmatrix} = [0 \quad *] \tag{4.25}$$

and let

$$W_2 := [w_{ij}^{(2)}] := W_1 J_5. \tag{4.26}$$

Let $J_6 = J(n, c_6, s_6) \in US_{2n}$ such that

$$[w^{(2)}_{2n-1,n}, w^{(2)}_{2n-1,2n}]\begin{bmatrix} c_6 & s_6 \\ -s_6 & c_6 \end{bmatrix} = [0 \quad *] \tag{4.27}$$

and let

$$W_3 = [w^{(3)}_{ij}] = W_2 J_6. \tag{4.28}$$

Let

$$J_7 = \left[\begin{array}{cc|cc} 1 & & & \\ & \ddots & & \\ & 1 & & \\ & & c_7 & s_7 \\ & & -s_7 & c_7 \\ \hline & & 1 & \\ & & & \ddots \\ & & & 1 \\ & & & & c_7 & s_7 \\ & & & & -s_7 & c_7 \end{array}\right] \in US_{2n}$$

such that

$$[w^{(3)}_{2n-1,2n-1}, w^{(3)}_{2n-1,2n}]\begin{bmatrix} c_7 & s_7 \\ -s_7 & c_7 \end{bmatrix} = \begin{bmatrix} 0 \\ * \end{bmatrix}. \tag{4.29}$$

Then,

$$W_4 = [w^{(4)}_{ij}] := \begin{bmatrix} W^{(4)}_{11} & W^{(4)}_{12} \\ W^{(4)}_{21} & W^{(4)}_{22} \end{bmatrix} = W_3 J_7 \triangleq \begin{bmatrix} \diagdown & \square \\ ** & \diagdown \end{bmatrix}. \tag{4.30}$$

$W_4^{-1} Y_5$ is still symplectic; thus it follows that $W^{(4)}_{21} W^{(4)*}_{22} = W^{(4)}_{22} W^{(4)*}_{21}$ and $W^{(4)*}_{22}$ has nonzero diagonal elements $w^{(4)}_{i,i}$, $i = n+1, \cdots, 2n-1$. Thus, it follows that $w^{(4)}_{2n,n-1} = 0$. Let $J_8 = J(n, c_8, s_8) \in US_{2n}$ such that

$$[w^{(4)}_{2n,n}, w^{(4)}_{2n,2n}]\begin{bmatrix} c_8 & s_8 \\ -s_8 & c_8 \end{bmatrix} = [0 \quad *]. \tag{4.31}$$

Then,

$$W_5 := \begin{bmatrix} W^{(5)}_{11} & W^{(5)}_{12} \\ W^{(5)}_{21} & W^{(5)}_{22} \end{bmatrix} := W_4 J_8 \triangleq \begin{bmatrix} \diagdown & \square \\ 0 & \diagdown \end{bmatrix}. \tag{4.32}$$

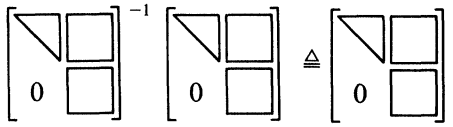
Now let $W^{(5)}_{11} = R_2 Z_2$ be an RQ-factorization of $W^{(5)}_{11}$; then

$$W_6 = W_5 \begin{bmatrix} Z_2^* & 0 \\ 0 & Z_2^* \end{bmatrix} \triangleq \begin{bmatrix} \diagdown & \square \\ 0 & \square \end{bmatrix} \tag{4.33}$$

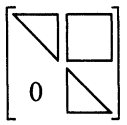
and

(4.34)

$$T := \begin{bmatrix} Z_2^* & 0 \\ 0 & Z_2^* \end{bmatrix} J_8^* J_7^* J_6^* J_5^* \begin{bmatrix} Z_1^* & 0 \\ 0 & Z_1^* \end{bmatrix} S \triangleq$$



is symplectic and therefore has the structure



and thus for

(4.35)
$$Q^* = \begin{bmatrix} Z_2^* & 0 \\ 0 & Z_2^* \end{bmatrix} J_8^* J_7^* J_6^* J_5^* \begin{bmatrix} Z_1^* & 0 \\ 0 & Z_1^* \end{bmatrix}$$

we have that

$$S = QT$$

is a QT-factorization as required.

Implicit single shift *DSQR*-step. Given a symplectic iterate M_i, we have to compute the factors Z_1, J_2, Z_2 from the QT-factorization and then produce the next iterate

(4.36)
$$M_{i+1} = \begin{bmatrix} Z_2 & 0 \\ 0 & Z_2 \end{bmatrix}^* J_2^* \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix}^* M_i \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix} J_2 \begin{bmatrix} Z_2 & 0 \\ 0 & Z_2 \end{bmatrix}.$$

Using Remark 3.25, we can easily obtain F, $H = \alpha e_n e_n^*$ and also C, $\tilde{M}$ from the given

$$M_i = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}.$$

Z_1 is then obtained by computing the QR-factorization of $\bar{s}F - I$ (or $\bar{s}M_{11} - I$, since they differ only in the last column). Then

(4.37)
$$M^{(1)} := \begin{bmatrix} Z_1^* & 0 \\ 0 & Z_1^* \end{bmatrix} M_i \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix} =: \begin{bmatrix} M_{11}^{(1)} & M_{12}^{(1)} \\ M_{21}^{(1)} & M_{22}^{(1)} \end{bmatrix} \triangleq$$

since Z_1 is upper Hessenberg by construction. J_1 is obtained from

$$Y_1 = \begin{bmatrix} Z_1^* & 0 \\ 0 & I \end{bmatrix} [\bar{s}\tilde{M} - C]$$

as described above. We use

(4.38)
$$y_{2n,n}^{(1)} = \bar{s}\tilde{m}_{2n,n} = \bar{s}\alpha,$$

and

(4.39)
$$y_{n,n}^{(1)} = e_n^* Z_1^* (\bar{s}\tilde{M}_{11} - I) e_n.$$

J_2 is obtained from $W^{(2)}_{2n,n}$, $W^{(2)}_{2n,2n}$, where

$$W_2 = J_1 W_1 = J_1 \begin{bmatrix} Z_1^* & 0 \\ 0 & I \end{bmatrix} (\tilde{M} - sC) \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix}$$

$$= J_1 \begin{bmatrix} Z_1^*(\tilde{M}_{11} - sI)Z_1 & Z_1^* \tilde{M}_{12} Z_1 \\ \tilde{M}_{21} Z_1 & (\tilde{M}_{22} - sF)Z_1 \end{bmatrix}$$

$$= J_1 \begin{bmatrix} M_{11}^{(1)} - sI & M_{12}^{(1)} \\ \alpha e_n e_n^* Z_1 & (I - sF^*)Z_1 \end{bmatrix}.$$

Thus

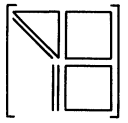
$$W^{(2)}_{2n,2n} = -s_1 e_n M_{12}^{(1)} e_n + c_1 e_n^*(I - sF^*) z_1 e_n \tag{4.40}$$

and

$$W^{(2)}_{2n,n} = -s_1 e_n^*(M_{11}^{(1)} - sI)e_n + c_1 \alpha e_n^* Z_1 e_n, \tag{4.41}$$

and J_2 is obtained by (4.15). Then

$$M^{(2)} := J_2^* M^{(1)} J_2 =: \begin{bmatrix} M_{11}^{(2)} & M_{12}^{(2)} \\ M_{21}^{(2)} & M_{22}^{(2)} \end{bmatrix} \triangleq \tag{4.42}$$



since J_2 does not change the structure of $M^{(1)}$. Using Algorithm 3.3 we compute $Z_2 \in U_n$ such that

$$M_{i+1} = \begin{bmatrix} Z_2^* & 0 \\ 0 & Z_2^* \end{bmatrix} M^{(2)} \begin{bmatrix} Z_2 & 0 \\ 0 & Z_2 \end{bmatrix} \tag{4.43}$$

is again an S-Hessenberg matrix.

Observe that due to the structure of $M^{(2)}$, we can perform Algorithm 3.3 with very thin Householder symplectic matrices

$$\begin{bmatrix} P_i & 0 \\ 0 & P_i \end{bmatrix},$$

where P_i is a usual Givens rotation in U_n, chasing only one subdiagonal element along the diagonal of $M_{11}^{(2)}$.

Implicit double shift *DSQR*-step for real M_i. Obtain again F, H from the iterate

$$M_i = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

as in Remark 3.25. Z_1 is obtained by performing an implicit double shift QR-step on $(\bar{s}F - I)(sF - I)$ using the usual Francis procedure, e.g., [13].

Let

$$M^{(1)} := \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1^* \end{bmatrix} M_i \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix} =: \begin{bmatrix} M_{11}^{(1)} & M_{12}^{(1)} \\ M_{22}^{(1)} & M_{22}^{(1)} \end{bmatrix}, \qquad F_1 = Z_1^* F Z_1. \tag{4.44}$$

Now by (4.13) we have

(4.45)

$$Y_1 = \begin{bmatrix} Z_1^* & 0 \\ 0 & I \end{bmatrix} L_2$$

$$= \begin{bmatrix} Z_1^*[(\bar{s}F-I)(sF-I)+|s|^2FGF^{-*}H] & |s|^2Z_1^*FGF^{-*}-Z_1^*G \\ |s|^2(F^*HF+F^*HGF^{-*}H+H)-2\,\mathrm{Re}\,(s)F^*H & |s|^2F^*HGF^{-*}+(\bar{s}I-F^*)(sI-F^*) \end{bmatrix}.$$

Thus the elements of Y_1 needed to compute J_1, J_2, J_3, J_4 are $y^{(1)}_{2n-1,n-1}$, $y^{(1)}_{2n,n-1}$, $y^{(1)}_{2n,n}$, $y^{(1)}_{2n-1,n}$, $y^{(1)}_{n-1,n-1}$, $y^{(1)}_{n,n}$, $y^{(1)}_{n-1,n}$.

Letting $F = [f_{ij}]$, $M_i = [m_{ij}]$, we then obtain the following formulae:

(4.46) $$y^{(1)}_{2n-1,n-1} = |s|^2\alpha f^2_{n,n-1},$$

(4.47) $$y^{(1)}_{2n,n-1} = |s|^2\alpha f_{n,n}f_{n,n-1},$$

(4.48) $$y^{(1)}_{2n,n} = |s|^2\alpha(f^2_{n,n}+1)+|s|^2\alpha^2 f_{n,n}m_{n,2n}-2\,\mathrm{Re}\,(s)\alpha f_{n,n},$$

(4.49) $$y^{(1)}_{2n-1,n} = |s|^2\alpha f_{n,n-1}f_{n,n}+|s|^2\alpha^2 f_{n,n-1}m_{n,2n}-2\alpha\,\mathrm{Re}\,(s)f_{n,n-1},$$

(4.50) $$y^{(1)}_{n-1,n-1} = e^*_{n-1}Z_1^*(\bar{s}F-I)(sF-I)e_{n-1},$$

(4.51) $$y^{(1)}_{n,n} = e^*_n Z_1^*(\bar{s}F-I)(sF-I)e_n+|s|^2\alpha e^*_n Z_1^*FM_{12}e_n,$$

(4.52) $$y^{(1)}_{n-1,n} = e^*_{n-1}(Z_1^*(\bar{s}F-I)(sF-I)+|s|^2\alpha Z_1^*FM_{12})e_n.$$

From these values we can then completely determine J_1, J_2, J_3, J_4 via the above described process. Then we have to determine J_5, J_6, J_7, J_8 from W_1 given by (4.21). We only need the 4 trailing 2×2 submatrices of the blocks of L_1, together with J_1, J_2, J_3, J_4 to compute those elements which are necessary to determine $J_5, \cdots, J_8$. Now let

(4.53)

$$L_3 := \begin{bmatrix} L^{(3)}_{11} & L^{(3)}_{12} \\ L^{(3)}_{21} & L^{(3)}_{22} \end{bmatrix} := \begin{bmatrix} Z_1^* & 0 \\ 0 & I \end{bmatrix} L_1 \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix}$$

$$=: [l_{ij}]$$

$$= \begin{bmatrix} (F_1-sI)(F_1-\bar{s}I)+F_1M^{(1)}_{12}M^{(1)}_{21} & F_1M^{(1)}_{12}-|s|^2M^{(1)}_{12}F_1 \\ M_{21}(F-2\,\mathrm{Re}\,(s)I+M_{12}H)Z_1+HZ_1 & [M_{21}M_{12}+(I-sF^*)(I-\bar{s}F^*)]Z_1 \end{bmatrix}.$$

Clearly the necessary submatrices

$$\begin{bmatrix} l_{n-1,n-1} & l_{n-1,n} \\ l_{n,n-1} & l_{n,n} \end{bmatrix}, \quad \begin{bmatrix} l_{n-1,2n-1} & l_{n-1,2n} \\ l_{n,2n-1} & l_{n,2n} \end{bmatrix}, \quad \begin{bmatrix} l_{2n-1,n-1} \\ l_{2n,n-1} \end{bmatrix}, \quad \begin{bmatrix} l_{2n-1,2n-1} & l_{2n-1,2n} \\ l_{2n,2n-1} & l_{2n,2n} \end{bmatrix}$$

are easily obtained. We omit the exact formulas here. Knowing these values of L_3 we produce the necessary submatrices of W_1

$$\begin{bmatrix} W_{2n-1,n-1} & W_{2n-1,n} \\ W_{2n,n-1} & W_{2n,n} \end{bmatrix}, \quad \begin{bmatrix} W_{2n-1,2n-1} & W_{2n-1,2n} \\ W_{2n,n-1} & W_{2n,2n} \end{bmatrix}$$

by multiplying $J_4J_3J_2J_1L_3$. The matrices J_5, J_6, J_7, J_8 are then obtained via (4.25), (4.27), (4.29), and (4.31).

Let

$$M^{(5)} = J_8^* J_7^* J_6^* J_5^* M^{(1)} J_5 J_6 J_7 J_8 =: \begin{bmatrix} M_{11}^{(5)} & M_{12}^{(5)} \\ M_{21}^{(5)} & M_{22}^{(5)} \end{bmatrix}. \tag{4.54}$$

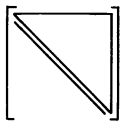
Now let

$$\begin{bmatrix} Z_2 & 0 \\ 0 & Z_2 \end{bmatrix} \in US_{2n}$$

such that

$$M_{i+1} := \begin{bmatrix} Z_2^* & 0 \\ 0 & Z_2^* \end{bmatrix} M^{(5)} \begin{bmatrix} Z_2 & 0 \\ 0 & Z_2 \end{bmatrix} \tag{4.55}$$

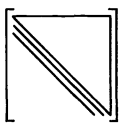
is again an S-Hessenberg matrix. Observe that by construction as in the QR-algorithm (e.g., [13])

$$F_1 = Z_1^* F Z_1 \triangleq$$ 

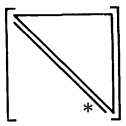
thus

$$M_{11}^{(1)} = Z_1^*(F + M_{12}\alpha e_n e_n^*) Z_1 = F_1 + Z_1^* M_{12} \alpha e_n e_n^*, \tag{4.56}$$

but since Z_1 is obtained from the QR-step applied to $(\bar{s}F - I)(sF - I)$ it follows that

$$Z_1 \triangleq$$  . (4.57)

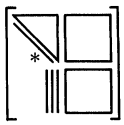
Thus

$$M_{11}^{(1)} \triangleq$$ 

is upper Hessenberg plus an extra entry in the last row

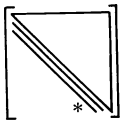
$$M_{21}^{(1)} = Z_1^* F^{-*} H Z_1 = F_1^{-*} Z_1^* H Z_1 \triangleq \begin{bmatrix} & 0 & \\ & & \end{bmatrix} \tag{4.58}$$

has only entries in the last three columns. $M_{12}^{(1)}$, $M_{22}^{(1)}$ are full in general. Thus

$$M_1^{(1)} \triangleq$$  . (4.59)

The transformations with J_5, J_6, J_7, J_8 do not change this structure. Thus, in the last step we transform $M^{(5)}$ to S-Hessenberg form using Algorithm 3.3.

Observe that after the first step of Algorithm 3.3, we have to transform the matrix in the (1, 1) block position, which is of the form



to upper Hessenberg form, which we do by chasing the bulge up along the diagonal, so we only need Householder reflections of length 4 for each row (e.g., [13]).

In both the single and the double shift algorithm, we monitor the effect of rounding errors on the rank 1 property of the (2, 1) block by computing the elements that have to be zero since it is rank 1, i.e., in the single shift algorithm the $(n-1)$th column and in the double shift algorithm columns $n-1$, $n-2$.

We do not give detailed descriptions of the implicit procedure for the *DSSZ* algorithm, first since they are more complicated and second since they involve unstable manipulations from the left. Since the *DSQR*-algorithm can be carried out stably even if B^{-1} does not exist, the *DSSZ*-algorithm is also always implicitly available by using the formulas in Remark 3.25.

Choice of shifts. For the iterates A_i, B_i, the matrix $K(A_i - \lambda B_i)K$ is upper Hessenberg. Thus, the Wilkinson shift [25] for $K(A_i - \lambda B_i)K$ would be to take eigenvalues from the 2×2 matrix in the top left corner of $I - \lambda F_i^*$, which are just the reciprocals of the eigenvalues of the 2×2 matrix in the top left corner of $A_i - \lambda B_i$ or M_i. If λ is a good estimate for an eigenvalue, then also $1/\bar{\lambda}$ is a good estimate and since we want the subspace corresponding to the eigenvalues of modulus $|\lambda| < 1$, we take shifts λ (or $\lambda, \bar{\lambda}$ in the double shift case) such that $|\lambda| > 1$. Using these shifts and our matrices $S_i = (M_i - \lambda I)^{-1}(M_i - \bar{\lambda} I)$, we essentially do steps of inverse iteration (e.g., [13] or [11]), and therefore have convergence in the top left corner of the two diagonal blocks.

Deflation. If during the iteration a subdiagonal element of $(M_i)_{11}$ becomes neglectably small compared to $\|F\| + \|G\| + \|H\|$, we set it to zero. If

$$M_i = \left[\begin{array}{cc|cc} M_{11} & M_{12} & M_{13} & M_{14} \\ 0 & M_{22} & M_{23} & M_{24} \\ \hline 0 & M_{32} & M_{33} & M_{34} \\ 0 & M_{42} & M_{42} & M_{44} \end{array}\right], \tag{4.60}$$

then by (3.28) also $M_{34} = 0$. Thus, we may then also set the corresponding whole block of $(M_i)_{22}$ to zero, and split the problem into the two subproblems of computing subspaces of

$$\begin{bmatrix} M_{11} & M_{13} \\ 0 & M_{33} \end{bmatrix}, \tag{4.61}$$

which is just a usual eigenvalue problem for M_{11} and M_{33} and can be treated by the QR algorithm for M_{11} and

$$\begin{bmatrix} M_{22} & M_{24} \\ M_{42} & M_{44} \end{bmatrix}, \tag{4.62}$$

which is of the same type as before.

Ordering of eigenvalues. For the optimal control problem we want to obtain all eigenvalues with modulus $|\lambda| < 1$ in the upper left corner of $A_i - \lambda B_i$ or M_i. Thus, possibly we have to exchange eigenvalues which, using Stewart's method [22], is essentially a QR-step with the eigenvalue (or pair of conjugate eigenvalues) as shifts applied to a 2×2 or 4×4 submatrix of $A_i - \lambda B_i$ or M_i.

Let Q, Z be chosen such that

$$Q(A-\lambda B)Z=\begin{bmatrix} T_{11} & 0 \\ 0 & I \end{bmatrix}-\lambda\begin{bmatrix} I & T_{12} \\ 0 & T_{11}^* \end{bmatrix} \tag{4.63}$$

or

$$Z^*MZ=\begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{11}^{-1} \end{bmatrix}:=[r_{ij}] \tag{4.64}$$

where T_{11}, R_{11} are upper triangular (or quasi-upper triangular in the real Schur form with complex eigenvalues). If a diagonal element r_{ii} of R_{11} is such that $|r_{ii}| > 1$, then let $D = J(i, c, s)$ such that

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}^*\begin{bmatrix} r_{i,i+n} \\ r_{i+n,i+n-r_{ii}} \end{bmatrix}=\begin{bmatrix} * \\ 0 \end{bmatrix}; \tag{4.65}$$

then in

$$D^*Z^*MZD \tag{4.66}$$

the diagonal elements r_{ii}, $r_{n+i,n+i}$ have been exchanged.

In the case of complex eigenvalues in the real Schur form, if the eigenvalues λ, $\bar{\lambda}$ of the 2×2 diagonal block

$$\begin{bmatrix} r_{i,i} & r_{i,i+1} \\ r_{i+1,i} & r_{i+1,i+1} \end{bmatrix}$$

in R_{11} have $|\lambda| > 1$, we perform a real double shift QR-step with the exact eigenvalues λ, $\bar{\lambda}$, to the submatrix

$$\begin{bmatrix} r_{i,i} & r_{i,i+1} & r_{i,i+n} & r_{i,i+n+1} \\ r_{i+1,i} & r_{i+1,i+1} & r_{i+1,i+n} & r_{i+n,i+n+1} \\ 0 & 0 & r_{i+n,i+n} & r_{i+n,i+n+1} \\ 0 & 0 & r_{i+n+1,i+n} & r_{i+n+1,i+n+1} \end{bmatrix}.$$

Computation of the deflating (invariant) subspace. In order to solve the optimal control problem we have to compute the deflating (invariant) subspace corresponding to the eigenvalues λ with $|\lambda| < 1$ of $\mathscr{A} - \lambda\mathscr{B}(\mathscr{M})$. There are essentially two ways to do this, which are exactly the same as in the continuous time (Hamiltonian) case (e.g., Byers [4]). Either we accumulate all the transformation matrices and then compute the deflating subspaces from the (1.1) and (2.1) blocks of the transformation, or we use the symmetric updating procedure of Byers and Mehrmann [6] to compute the solution X directly. For a comparison of these two procedures see [6].

5. Conclusions. Comparing the work needed for the reduction to S-Hessenberg form and per iteration step, we get the following approximate flop counts. (A flop is defined to be the work of evaluating the FORTRAN statement $A[I, J] = A[I, J] - S^*A[I, K]$; see Table 5.1.)

TABLE 5.1
Approximate flop counts.

	Flops
Algorithm 3.3 for full symplectic pencil	$10n^3$
Algorithm 3.3 for full symplectic matrix	$\frac{15}{3}n^3$
Accumulating UV	$\frac{2}{3}n^3$
One implicit single shift $DSQR$-step	$40n^2$
Accumulating transformations	$16n^2$
One implicit double shift $DSQR$-step	$62n^2$
Accumulating transformations	$24n^2$
Reduction to Hessenberg form for arbitrary $2n \times 2n$ matrix	$\frac{40}{3}n^3$
Accumulating transformations	$\frac{16}{3}n^3$
Reduction to invariant form in QZ-algorithm	$40n^3$
Accumulating transformations in Z	$12n^3$
One implicit double shift QR-step	$24n^2$
Accumulating transformations	$24n^2$
One implicit QZ-step	$52n^2$
Accumulating transformations	$32n^2$

Numbers for QR and QZ are taken from [13].

Due to the special eigenstructure, which is preserved by the algorithm, we have (and may force) convergence of always at least two or even four eigenvalues at a time. Due to the loss in structure this is not quite the case in the QR- or QZ-algorithm. The work required by the described $DSQR$-algorithm is, between that of the QZ- and the QR-algorithm, since the S-Hessenberg form is not as thin as the Hessenberg form. In many cases we cannot explicitly produce $\mathscr{B}^{-1}$ without causing large roundoff errors; thus the direct application of the QR-algorithm is not advisable. The described algorithm is faster than the general QZ and in any case is, since it has the symplectic structure, more advisable from the point of view that any internal structure should be (if possible) preserved for stability reasons.

We have shown that it is possible to produce a numerically stable algorithm for the single input/output discrete linear quadratic control problem. The principle of Byers method for the continuous case can therefore also be applied in the discrete case.

Acknowledgments. We thank A. Bunse-Gerstner and L. Elsner for many helpful discussions.

REFERENCES

[1] W. F. ARNOLD III, *Numerical solution of algebraic Riccati equations*, Technical Report NWCTP 65L, Naval Weapons Research Center China Lake, 1984.

[2] W. F. ARNOLD III AND A. J. LAUB, *Generalized eigenproblem algorithm and software for algebraic Riccati equations*, Proc. IEEE, 72 (1984), pp. 1746–1754.

[3] A. J. BENDER AND A. J. LAUB, *The linear quadratic optimal regulator for descriptor systems: discrete-time case*, Automatica, 1986, to appear.
[4] R. BYERS, *Hamiltonian and symplectic algorithms for the algebraic Riccati equation*, Ph.D. thesis, Cornell University, Ithaca, NY, January 1983.
[5] ———, *A Hamiltonian QR-algorithm*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 212–229.
[6] R. BYERS AND V. MEHRMANN, *Symmetric updating of the solution of the algebraic Riccati equation*, in Proc. X Symposium on Operations Research 1985, Methods of Operations Research 54, Beckmann, Gaede, Ritter, Schneeweis, eds., 1986, pp. 117–125.
[7] A. BUNSE-GERSTNER, *Matrix factorization for symplectic QR-like methods*, Linear Algebra Appl., 83 (1986), pp. 49–77.
[8] ———, *QR-like algorithms*, Habilitationsschrift, Universität Bielefeld, July 1986.
[9] ———, *Eigenvalue algorithms for matrices with special structure*, Colloquia Math. Soc. János Bolyai, Numerical Methods, Miskolc, Hungary, 1986, to appear.
[10] 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.
[11] W. BUNSE AND A. BUNSE-GERSTNER, *Numerische Lineare Algebra*, Teubner-Verlag, Stuttgart, 1985.
[12] S. L. CAMPBELL, *Singular Systems of Differential Equations*, Pitman, San Francisco, CA, 1980.
[13] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, North Oxford Academic, Oxford, 1983.
[14] A. J. LAUB AND K. MEYER, *Canonical forms for symplectic and Hamiltonian matrices*, Celestial Mech., 9 (1974), pp. 213–238.
[15] A. J. LAUB, *Schur techniques for Riccati differential equations*, in Feedback Control of Linear and Nonlinear Systems, D. Hinrichsen and A. Isidori, eds., Springer-Verlag, New York, Berlin, 1982, pp. 165–174.
[16] C. B. MOLER AND G. W. STEWART, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241–256.
[17] C. PAIGE AND C. F. VAN LOAN, *A Schur decomposition for Hamiltonian matrices*, Linear Algebra Appl., 41 (1981), pp. 11–32.
[18] T. PAPPAS, A. J. LAUB, AND N. R. SANDELL, *On the numerical solution of the discrete-time algebraic Riccati equation*, IEEE Trans. Autom. Control, AC-25 (1980), pp. 631–641.
[19] A. P. SAGE, *Optimum Systems Control*, Prentice-Hall, Englewood Cliffs, NJ, 1966.
[20] G. W. STEWART, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
[21] ———, *Introduction to Matrix Computation*, Academic Press, New York, 1973.
[22] ———, HQR3 *and* EXCHNG: FORTRAN *subroutines for calculating and ordering the eigenvalues of a real upper Hessenberg matrix*, ACM Trans. Math. Software, 2 (1975), pp. 275–280.
[23] J. STOER AND R. BULIRSCH, *Einführung in die Numerische Mathematik* II, Springer-Verlag, New York, Berlin, 1973.
[24] P. VAN DOOREN, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.
[25] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

010

A DETERMINANT IDENTITY AND ITS APPLICATION IN EVALUATING FREQUENCY RESPONSE MATRICES*

P. MISRA† AND R. V. PATEL†

Abstract. This paper is concerned with the computation of frequency response matrices of linear multivariable systems described by their state-space equations. A determinant identity is used to evaluate these matrices that play an important role in frequency domain analysis and design of linear multivariable systems. The algorithm proposed here is believed to be considerably faster and at least as accurate as other existing ones. This is illustrated by means of operations counts and numerical examples. It is also shown that the proposed method can be easily adapted for implementation in a parallel processing environment.

Key words. frequency response, computational methods, linear systems

AMS(MOS) subject classifications. 93B40, 93C05

1. Introduction. Many so-called classical and modern control system design methods for linear time-invariant systems (e.g., see [1]–[6]) use the frequency response characteristics of a system to design controllers which achieve desired stability and robustness properties for the resulting closed-loop systems. Hence, an efficient and accurate computation of frequency response matrices is of considerable importance. In this paper, we consider the linear time-invariant, multivariable system described by

$$\dot{\underline{x}}(t) = A\underline{x}(t) + B\underline{u}(t), \tag{1.1a}$$

$$\underline{y}(t) = C\underline{x}(t) + D\underline{u}(t) \tag{1.1b}$$

where $\underline{x}(t) \in \mathbb{R}^n$, $\underline{u}(t) \in \mathbb{R}^m$, and $\underline{y}(t) \in \mathbb{R}^p$. The frequency response matrix $W(j\omega)$ of the system (A, B, C, D) is given by

$$W(j\omega) = C(j\omega I_n - A)^{-1}B + D. \tag{1.2}$$

Computation of frequency response usually requires evaluation of $W(j\omega_k)$ at a large number of frequencies ω_k, $k = 1, \cdots, N$. If the system description is given in terms of the transfer function matrix $W(s)$, then the computation of frequency response is a relatively simple matter. However, if the state-space description (A, B, C, D) is given, then the problem is not so straightforward computationally. Obtaining the frequency response by first converting the state-space description to a transfer function description is justifiable only when the initial cost of computing the transfer function matrix is offset by the number of frequencies at which the frequency response is desired. From the operations count in § 4, it is possible to determine approximately when a direct determination of frequency response would be more economical than computing a transfer function matrix followed by evaluating this matrix at various frequencies. In [8], a method for computing frequency response was proposed which, starting from a given state-space description, determines the frequency response matrix by first reducing the state matrix to an upper Hessenberg matrix and then solving a system of n simultaneous linear equations. Depending on the number of frequencies at which the frequency response matrix is desired,

* Received by the editors May 12, 1987; accepted for publication October 1, 1987. This research was supported by the Natural Sciences and Engineering Research Council of Canada under grant A1345. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† Department of Electrical Engineering, Concordia University, Montreal, Canada H3G 1M8.

the methods described in subsequent sections are comparable to or more efficient than this or other existing methods.

The layout of this paper is as follows. In § 2, we introduce some background material from linear algebra and control theory that forms the basis of the algorithms. Section 3 describes an algorithm for the computation of frequency response matrices of a given system and discusses its properties. Section 4 discusses computational requirements using various existing methods for evaluating the frequency response matrices and illustrates the numerical performance of the algorithm by means of some examples.

2. Preliminary considerations. We shall use the following facts from linear algebra and control theory for the development of the algorithms.

FACT 1. A single-input system $(A, \underline{b})$ can always be reduced to an *upper Hessenberg form* (UHF) [7], [9], by means of an orthogonal similarity transformation matrix T such that $F = T^T A T$ is an upper Hessenberg matrix and $\underline{g} = T^T \underline{b} = [g_0 \cdots 0]^T$. The element $g_1 \neq 0$ and F is an *unreduced* upper Hessenberg matrix if and only if $(A, \underline{b})$ is a controllable pair.

Further, if the system is not completely controllable, then the above transformation will reduce $(A, \underline{b})$ to $(F, \underline{g})$ such that

$$F = \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix} \tag{2.1a}$$

and

$$\underline{g} = \begin{bmatrix} \underline{g}_1 \\ 0 \end{bmatrix} \tag{2.1b}$$

where $F_{11} \in \mathbb{R}^{n_c \times n_c}$ is an unreduced upper Hessenberg matrix, $F_{22} \in \mathbb{R}^{(n-n_c)\times(n-n_c)}$ and $\underline{g}_1$ $(\in \mathbb{R}^{n_c}) = [g_1 0 \cdots 0]$. Note that $(F_{11}, \underline{g}_1)$ is a controllable pair and the eigenvalues of F_{22} correspond to the uncontrollable modes of $(A, \underline{b})$.

FACT 2. Similar results can be stated for a single-output system $(A, \underline{c}^T)$, except that the results apply to the observability properties of the system with $F = T^T A T$ in UHF and $\underline{c}^T T = [0 \cdots 0 c_n]$. The unobservable modes of the system are then defined in a similar manner.

FACT 3. For a single-input, single-output system $(A, \underline{b}, \underline{c}^T)$, we may write [5],

$$\det(j\omega I_n - A + \underline{b}\underline{c}^T) = \det(j\omega I_n - A) + \underline{c}^T \operatorname{adj}(j\omega I_n - A)\underline{b}.$$

Further,

$$\begin{aligned} \underline{c}^T (j\omega I_n - A)^{-1} \underline{b} &= \frac{\underline{c}^T \operatorname{adj}(j\omega I_n - A)\underline{b}}{\det(j\omega I_n - A)} \\ &= \frac{\det(j\omega I_n - A + \underline{b}\underline{c}^T)}{\det(j\omega I_n - A)} - 1. \end{aligned} \tag{2.2}$$

Now, in (1.2), the (i, l)th element of $W(j\omega)$ is given by

$$w_{il}(j\omega) = \underline{c}_i^T (j\omega I_n - A)^{-1} \underline{b}_l + d_{il} \tag{2.3}$$

which may be written, using (2.2) as

$$w_{il}(j\omega) = \frac{\det(j\omega I_n - A + \underline{b}_l \underline{c}_i^T)}{\det(j\omega I_n - A)} - 1 + d_{il}. \tag{2.4}$$

FACT 4. The determinant of a matrix whose kth column can be expressed as a sum of column vectors $\underline{a}_k + \underline{\hat{a}}_k$, may be written as [11]

$$\det(\underline{a}_1 \underline{a}_2 \cdots \underline{a}_k + \underline{\hat{a}}_k \cdots \underline{a}_n) = \det(\underline{a}_1 \underline{a}_2 \cdots \underline{a}_k \cdots \underline{a}_n) + \det(\underline{a}_1 \underline{a}_2 \cdots \underline{\hat{a}}_k \cdots \underline{a}_n). \tag{2.5}$$

FACT 5. An upper Hessenberg matrix $A \in \mathbb{R}^{n \times n}$ can be factored into the product of a *unit lower bidiagonal* matrix L and an *upper triangular* matrix U [10]–[11]. This decomposition is called an LU decomposition of A. The determinant of A is given by $\prod_{i=1}^{n} u_{ii}$, where u_{ii} denotes the ith diagonal element of U.

3. Computation of frequency response matrices. The method for computing the frequency response matrices outlined in this section determines one row of the matrix at a time, i.e., we evaluate the frequency response of the multi-input, single output systems $(A, B, \underline{c}_i^T)$, $i = 1, \cdots, p$. Note that for the sake of clarity, we have dropped the matrix D from the system description. It can be easily incorporated in the final frequency response matrices by a simple addition. In evaluating the frequency response, each triple described above is first reduced to the condensed form described in Fact 2. This reduction is done only once for a given set of frequencies ω_k, $k = 1, \cdots, N$.

Equation (2.4) can be rewritten as

$$w_{il}(j\omega) = \frac{\det(j\omega I_n - A + \underline{b}_l \underline{c}_i^T) - \det(j\omega I_n - A)}{\det(j\omega I_n - A)} \tag{3.1}$$

$$= \frac{\det(\underline{\tilde{a}}_1, \cdots, \underline{\tilde{a}}_{n-1}, \underline{\tilde{a}}_n + \underline{b}_l c_{in}) - \det(\tilde{A})}{\det(\tilde{A})} \tag{3.2}$$

where $\tilde{A} = (j\omega I_n - A)$, $\underline{\tilde{a}}_k$ is the kth column of $\tilde{A}$, and c_{in} is the nth element of $\underline{c}_i$ and is the only nonzero element of the ith row of the output matrix C. Then, using the determinant identity in Fact 4, (3.2) may be simplified to

$$w_{il}(j\omega) = \det(\bar{A})/\det(\tilde{A}) \tag{3.3}$$

where $\bar{A} = (\underline{\tilde{a}}_1, \cdots, \underline{\tilde{a}}_{n-1}, \underline{b}_l c_{in})$. Note that $\bar{A}$ differs from $\tilde{A}$ in its last column only. From the LU decomposition of $\tilde{A} = \tilde{L}\tilde{U}$, we have

$$\det(\tilde{A}) = \det(\tilde{U}) = \prod_{r=1}^{n} \tilde{u}_{rr}. \tag{3.4}$$

Moreover, changing the last column of $\tilde{A}$ affects only the last column of $\tilde{U}$. Therefore (3.4) gives

$$w_{il}(j\omega) = \frac{\bar{n}_{il}}{\tilde{d}} = \frac{\bar{u}_{nn} \prod_{r=1}^{n-1} \tilde{u}_{rr}}{\tilde{u}_{nn} \prod_{r=1}^{n-1} \tilde{u}_{rr}} = \frac{\bar{u}_{nn}}{\tilde{u}_{nn}}. \tag{3.5}$$

In (3.5) above, $\tilde{u}_{nn}$ is the (n, n)th element of the matrix $\tilde{U}$ and $\bar{u}_{nn}$ is the (n, n)th element of the matrix $\bar{U}$ in the LU decomposition of $\bar{A}$. Therefore, instead of computing the determinants in (3.3), we only need to find the ratio in (3.5). A small saving in computation can be achieved by noting that $\bar{u}_{nn} = \hat{u}_{nn} c_{in}$, where $\hat{u}_{nn}$ is the (n, n)th element of the matrix $\hat{U}$ in the LU decomposition of the matrix $\hat{A} = [\underline{\tilde{a}}_1, \cdots, \underline{\tilde{a}}_{n-1}, \underline{b}_l]$. The (i, l)th element of the frequency response matrix is, therefore, given by

$$w_{il}(j\omega) = \frac{\hat{u}_{nn}}{\tilde{u}_{nn}} c_{in}. \tag{3.6}$$

3.1. An algorithm for computing the frequency response matrices. Assume that the triples in their condensed form described by Fact 2 are denoted by $(A^{(i)}, B^{(i)}, \underline{c}_i^T)$. Then, the algorithm based on the above discussion may be formally given as follows.

ALGORITHM 3.1.

for $i = 1{:}p$,
  reduce the ith triple to the condensed form $(F^{(i)}, G^{(i)}, \underline{h}_i^T) :=$
  $(T_i^T A T_i, T_i^T B, \underline{c}_i^T T_i)$;
  set the observable subsystem to $(A^{(i)}, B^{(i)}, \underline{c}_i^T) := O(F^{(i)}, G^{(i)}, \underline{h}_i^T)$
  set $n :=$ dimension of the observable subsystem;
  for $k = 1{:}N$,
    evaluate $\tilde{u}_{nn}^{(k)} :=$ (n, n)th element of $\tilde{U}^{(k)}$ in the LU decomposition of $(j\omega_k I - A^{(i)})$;
    for $l = 1{:}m$,
      $\hat{A}^{(i)} := (j\omega_k I - A^{(i)})$ with last column replaced by $\underline{b}_l^{(i)}$
      evaluate $\hat{u}_{nn}^{(k,l)} :=$ (n, n)th element of $\hat{U}^{(k,l)}$ in the LU decomposition of $\hat{A}^{(i)}$
      $$g_{il}(j\omega_k) = \frac{\hat{u}_{nn}^{(k,l)}}{\tilde{u}_{nn}^{(k)}} c_{in};$$
    end;
  end;
end;

At the end of the algorithm, we get the required frequency response matrix $W(j\omega)$ at N desired values of ω_k.

3.2. Remarks about the algorithm. (1) The triples $(A^{(i)}, B^{(i)}, \underline{c}_i^T)$ are in the special condensed form described in § 2. As shown in that section, only the last element of $\underline{c}_i^T$ is nonzero. As a result, forming $(j\omega_k I_n - A^{(i)} + \underline{b}_l^{(i)} \underline{c}_i^T)$ retains the upper Hessenberg structure of $A^{(i)}$. Also, the matrix $\hat{U}^{(k,l)}$ differs from $\tilde{U}^{(k)}$ in only its last column. This enables us to compute the frequency response matrix without actually calculating the determinants in (3.3), thereby reducing the number of operations.

(2) The LU decomposition of the upper Hessenberg matrix $\hat{A}$ requires only $\frac{1}{2}n^2$ floating point operations. Once the lower subdiagonal of $\hat{L}$ is known, subsequent evaluations of $\hat{u}_{nn}^{(k,l)}$ for all $\underline{b}_l$, $l = 1, \cdots, m$, require only nm extra operations.

(3) An error analysis of the LU decomposition of a matrix A with L being a unit lower bidiagonal matrix yields

$$LU = A + E$$

where L and U are exact matrices for a slightly perturbed matrix A. The elements of the error matrix E satisfy [10], [11]

$$|E_{ij}| \leqq n\pi\beta\gamma 10^{-t}$$

where n is the order of the matrix, π is some constant of order unity, β is the largest element of the matrix A, and $\gamma \leqq 2^{n-1}$. Although 2^{n-1} appears to be a rapidly growing function, in practice for upper Hessenberg matrices, large growth factors γ are almost never encountered. Moreover, if the inner products in the LU decomposition are accumulated in double precision, the factor n also disappears. The discussion above does not permit us to make a strong statement about the stability of the algorithm, but for all

practical purposes, the results obtained from using the proposed algorithm will, in general, be very reliable.

(4) The problem of evaluating the frequency response matrix can be divided into p (= the number of outputs) independent subproblems. Therefore, p processors may be employed to compute the frequency response. This will reduce the actual time of computation significantly. If, however, only one processor is used and if the number of inputs is smaller than the number of outputs, then computing the frequency response of the dual system would enable further reduction in the computational effort. This will become clear from the operations count given later.

(5) Several excellent techniques for solution of sparse simultaneous linear equations exist in the literature. Therefore, if the state matrix of the given system is sparse, it may be advantageous to consider the original system instead of its condensed form. Moreover, if a multiprocessor or an array processor is used, we can employ techniques for parallel solution of simultaneous linear equations [12].

(6) It is worth mentioning that any technique for efficient evaluation of the determinants of Hessenberg matrices may be used to determine the frequency response, e.g., Hyman's method and its variations [10], [13]. But such methods may run into floating point overflow/underflow as the value of ω increases. The proposed method (as well as the method in [8]) do not suffer from this drawback.

(7) The algorithm proposed above uses complex arithmetic. The use of complex arithmetic can be avoided by making minor modifications to the algorithm. The real and imaginary parts can be computed independently as described below.

Consider the given triple (A, B, C); its frequency response matrix can be obtained by solving

$$(j\omega I - A)Z = B \tag{3.7}$$

for Z and then computing

$$G(j\omega) = CZ. \tag{3.8}$$

Let $Z = Z_1 + jZ_2$; then equating the real and imaginary parts on both sides of (3.7), we get

$$\begin{bmatrix} -A & -\omega I \\ \omega I & -A \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} B \\ 0 \end{bmatrix}. \tag{3.9}$$

It is easy to see that in (3.9)

$$Z_1 = \frac{1}{\omega} AZ_2 \tag{3.10}$$

and

$$Z_2 = -\omega I(\omega^2 I + A^2)^{-1} B. \tag{3.11}$$

The term $\hat{G}(\omega^2) = (\omega^2 I + A^2)^{-1} B$ can be evaluated by applying Algorithm 3.1 to the system $(-A^2, B, I)$. Then, $Z_1 = -A\hat{G}(\omega^2)$ and $Z_2 = -\omega\hat{G}(\omega^2)$. Note that only real arithmetic is used in computing both Z_1 and Z_2. The frequency response is then given by $CZ_1 + jCZ_2$. Although the above approach uses only real arithmetic, it is important to note that in forming the matrix A^2, a significant amount of information may be lost for ill-conditioned systems.

(8) In the transformed triples $(A^{(i)}, B^{(i)}, \underline{c}_i^T)$, the matrix $A^{(i)}$ is an unreduced upper Hessenberg matrix and $c_{in} \neq 0$ if and only if A is completely observable from the ith

output. However, if that is not the case, then $A^{(i)}$ will have a block upper triangular structure and the system equations may be rewritten as

$$\begin{bmatrix} \dot{\underline{x}}_1 \\ \dot{\underline{x}}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \underline{u},$$

$$y_i = [0 \quad \underline{c}_i^T] \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \end{bmatrix}$$

where A_{22} is an unreduced upper Hessenberg matrix and $\underline{c}_i^T = [0\,0 \cdots c_{in}]$. The observable subsystem is $(A_{22}, B_2, \underline{c}_i^T)$ and the frequency response is given by

$$W(j\omega) = \underline{c}_i^T (j\omega I - A_{22})^{-1} B_2.$$

Since the system being considered now has an order equal to the dimension of A_{22}, the computational effort is accordingly reduced.

4. Computational considerations and numerical examples. In this section, we will first compare the computational requirements for various existing methods for evaluating frequency response matrices and then illustrate the accuracy of the proposed method by applying it to determine several frequency response matrices for an *ill-conditioned* system.

4.1. Operations count. Here we compare the operations count for various efficient methods for computing frequency response matrices. We consider three methods: (1) the method in [8], (2) the method proposed in the previous section, and (3) the method of first computing the transfer function matrix and then evaluating it at various desired frequencies.

Method in [8]. In this method, matrix A is transformed to an upper Hessenberg matrix while matrices B and C have no specific structure. An LU decomposition of $(j\omega I - A)$ is carried out and Z is obtained from $UZ = L^{-1}B$, where U and L are, respectively, upper triangular and unit lower bidiagonal matrices. The frequency response for one value of ω is then given by $W(j\omega) = CZ$. When efficiently implemented, the above steps together with an initial reduction of A to an upper Hessenberg form and corresponding transformations on B and C require approximately $(5/3)n^3 + (m + p)n^2$ (real) and $(\frac{1}{2})[(p + 1)n^2 + 2nmp]N$ (complex) operations for N values of ω.

Proposed method. The proposed method requires an initial reduction of several multi-input, single-output systems to a condensed form. This reduction requires approximately $(5/3)(n + m + 1)n^2p$ (real) operations. For each value of frequency, evaluation of $\tilde{u}_{nn}^{(i)}$ in Algorithm 3.1 requires $(\frac{1}{2})n^2$ operations and subsequent m values of $\tilde{u}_{nn}^{(i)}$ for all inputs require a total of nm operations. This is done for each triple $(A, B, \underline{c}_i^T)$. Therefore, $W(j\omega)$ can be evaluated in approximately $(5/3)(m + n + 1)n^2p$ (real) and $(p/2)(n^2 + 2nm)$ N (complex) floating point operations. Further saving can be achieved by considering the dual system if $p > m$, as can be easily seen from the expression above.

Considering remark (7), we note that the operations count given for Algorithm 3.1 above corresponds to the case when the system is observable from each of the outputs. However, this is usually not the case when very high order systems are considered. If a system is not observable from the ith output, the frequency response calculations are carried out on a lower order subsystem and a significant saving in the computational effort can be achieved. To illustrate the above point, consider a 40th order system with 5 inputs and 5 outputs and 100 values of frequency. If each of the outputs can observe only 20 states, the method in [8] requires approximately 2,440,000 "flops" (floating point operations) compared to 1,213,000 flops required by the proposed method.

Evaluating the transfer function matrix. Evaluating a transfer function matrix using the method proposed in [14] requires approximately $(5/3)(n^3 + n_c^3)m + 8(n^2 + n_c^2)mp + ((1/6)n_{co}^3 + n_{co}^2)mp$ flops. The scalars n_c and n_{co} correspond to the dimensions of controllable and controllable as well as observable subsystems, respectively. Assume that in the example being considered above, only 20 states are controllable from each of the inputs. Further, let only 15 states be observable from each output. Then, evaluating the transfer function matrix will require approximately 560,000 flops. Evaluation of frequency response matrices for 100 different values of ω will require approximately $n_{co}^2 mpN$ more flops. For the example under consideration, this is approximately 540,000 flops, giving a total of 1,100,000 flops. Note that this approach becomes extremely efficient if the number N is very large, because once the transfer function is known, it requires a very small number of computations for evaluating the frequency response matrix for different values of ω.

The above operations counts are approximate figures given for comparison. In practice, frequency responses may be computed with slightly less or more computational effort, depending on the controllability and observability properties of the system under consideration.

4.2. Numerical example. For the purpose of illustrating the accuracy of the proposed method, we will consider an extremely *ill-conditioned,* 9th order boiler model [15]. The frequency response was first calculated in double precision using the proposed method as well as the method in [8]. The results agreed up to the 15th significant digit. For the sake of comparison, we shall call the frequency response calculated in double precision as the "actual" response. Next, the frequency response matrix was obtained in single precision, using (1) the proposed method, (2) by first evaluating the transfer function matrix, and (3) using the method in [8], for a selected number of frequencies. The results for the three methods are shown in Tables 4.1–4.4 for the (1, 1) element of the frequency response matrix. The underlined digits indicate the accuracy of results for the specified frequencies using the three methods.

TABLE 4.1
(1,1) *element of frequency response matrix for* $\omega = 1$.

Actual	$-1.764752693176270d + 02 + 7.363956117630005d + 01i$
Proposed	$-1.764752702713013d + 02 + 7.363956117330102d + 01i$
Method [8]	$-1.764752664566040d + 02 + 7.363956403732300d + 01i$
Method [14]	$-1.764752655029297d + 02 + 7.363956165313721d + 01i$

TABLE 4.2
(1,1) *element of frequency response matrix for* $\omega = 10$.

Actual	$-2.125151613822383d + 00 + 6.456438212270109d - 02i$
Proposed	$-2.125151604413986d + 00 + 6.456437520682812d - 02i$
Method [8]	$-2.125151515007019d + 00 + 6.456438917666674d - 02i$
Method [14]	$-2.125151574611664d + 00 + 6.456438358873129d - 02i$

TABLE 4.3
(1,1) *element of frequency response matrix for* $\omega = 100$.

Actual	$-2.097382268402725d - 02 + 4.376591209620528d - 05i$
Proposed	$-2.097382280044258d - 02 + 4.376592733024154d - 05i$
Method [8]	$-2.097382198553532d - 02 + 4.376593096822035d - 05i$
Method [14]	$-2.097382233478129d - 02 + 4.376591778054717d - 05i$

TABLE 4.4
(1,1) *element of frequency response matrix for* $\omega = 1000$.

Actual	$-2.096995310680062d-04+4.350784624412990d-08i$
Proposed	$-2.096995312967920d-04+4.350783444628803d-08i$
Method [8]	$-2.096995267493185d-04+4.349920956769893d-08i$
Method [14]	$-2.096995276588132d-04+4.350785731688234d-08i$

5. Conclusions. Using a determinant identity, a computationally efficient method for determining the frequency response matrices of linear multivariable systems given in state space form was presented. The properties and performance of the proposed method were compared with those of existing methods.

REFERENCES

[1] I. HOROWITZ, *Synthesis of Feedback Systems*, Academic Press, New York, 1963.

[2] H. H. ROSENBROCK, *Computer-Aided Control Systems Design*, Academic Press, London, 1974.

[3] I. POSTLETHWAITE AND A. G. J. MACFARLANE, *Complex Variable Methods for Linear Multivariable Feedback Systems*, Taylor & Francis, London, 1980.

[4] D. H. OWENS, *Multivariable root-loci: an emerging design tool*, IEEE Internat. Conference on Control, Warwick, 1981.

[5] R. V. PATEL AND N. MUNRO, *Multivariable System Theory and Design*, Pergamon Press, Oxford, 1982.

[6] M. VIDYASAGAR, *Control Systems Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

[7] P. VAN DOOREN AND M. VERHAEGAN, *The use of condensed forms in linear systems theory*, AMS-IMS-SIAM 1984 Joint Summer Research Conference on Linear Algebra and Its Role in System Theory, Maine, July 29–August 4, 1984.

[8.] A. J. LAUB, *Efficient multivariable frequency response computations*, IEEE Trans. Automat. Control, 26 (1981), pp. 407–408.

[9] P. MISRA AND R. V. PATEL, *A computational method for frequency response of multivariable systems*, Proc. 24th IEEE Conference on Decision and Control, Ft. Lauderdale, FL, 1985, pp. 1248–1249.

[10] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, Oxford, 1965.

[11] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[12] R. H. BARLOW AND D. J. EVANS, *Parallel algorithms for iterative solution to linear systems*, Computer J., 25 (1982), pp. 56–60.

[13] M. A. HYMAN, *Eigenvalues and eigenvectors of general matrices*, 12th National Meeting of A.C.M., Houston, TX, 1957.

[14] P. MISRA AND R. V. PATEL, *Computation of transfer function matrices of linear multivariable systems*, Automatica, 23 (1987), pp. 635–640.

[15] E. J. DAVISON AND S. H. WANG, *Properties and calculation of transmission zeros of linear multivariable systems*, Automatica, 10 (1974), p. 643.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

011

ON MINIMIZING THE MAXIMUM EIGENVALUE OF A SYMMETRIC MATRIX*

MICHAEL L. OVERTON†

Abstract. An important optimization problem that arises in control is to minimize $\varphi(x)$, the largest eigenvalue (in magnitude) of a symmetric matrix function of x. If the matrix function is affine, $\varphi(x)$ is convex. However, $\varphi(x)$ is not differentiable, since the eigenvalues are not differentiable at points where they coalesce. In this paper an algorithm that converges to the minimum of $\varphi(x)$ at a quadratic rate is outlined. Second derivatives are not required to obtain quadratic convergence in cases where the solution is strongly unique. An important feature of the algorithm is the ability to split a multiple eigenvalue, if necessary, to obtain a descent direction. In these respects the new algorithm represents a significant improvement on the first-order methods of Polak and Wardi and of Doyle. The new method has much in common with the recent work of Fletcher on semidefinite constraints and Friedland, Nocedal, and Overton on inverse eigenvalue problems. Numerical examples are presented.

Key words. nonsmooth optimization, nondifferentiable optimization, convex programming, semidefinite constraints, minimizing maximum singular value

AMS(MOS) subject classifications. 65F99, 65K10, 90C25

1. Introduction.

Many important optimization problems involve eigenvalue constraints. For example, in structural engineering we may wish to minimize the cost of some structure subject to constraints on its natural frequencies. A particularly common problem, which arises in control engineering, is

$$\min_{x \in \mathbb{R}^m} \varphi(x) \tag{1.1}$$

where

$$\varphi(x) = \max_{1 \leq i \leq n} |\lambda_i(A(x))|, \tag{1.2}$$

$A(x)$ is a real symmetric $n \times n$ matrix-valued affine function of x, and

$$\{\lambda_i(A(x)), i = 1, \cdots, n\}$$

are its eigenvalues. Since $A(x)$ is an affine function, it may be written

$$A(x) = A_0 + \sum_{k=1}^{m} x_k A_k.$$

The function $\varphi(x)$ is convex, since the largest eigenvalue of a matrix is a convex function of the matrix elements. An important special case is

$$A_k = e_k e_k^T \tag{1.3}$$

* Received by the editors February 1, 1987; accepted for publication October 1, 1987. This work was supported in part by the National Science Foundation under grant DCR-85-02014. Some of the computer program development was performed at Stanford University, Stanford, California with support from the Office of Naval Research under contract ONR N00014-82-K-0335. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. This work was completed while the author was on sabbatical leave at the Centre for Mathematical Analysis and Mathematical Sciences Research Institute, Australian National University, Canberra, Australia.

where e_k is the kth column of the identity matrix so that

$$A(x) = A_0 + \text{Diag}\,(x). \tag{1.4}$$

Note that the problem of minimizing the maximum singular value of a nonsymmetric matrix-valued affine function $G(x)$ may be written in the form (1.1) since the eigenvalues of

$$\begin{bmatrix} 0 & G(x) \\ G(x)^T & 0 \end{bmatrix}$$

are (plus and minus) the singular values of $G(x)$. (Undoubtedly savings could be gained by treating the singular value problem more directly.)

The difficulty in minimizing $\varphi(x)$ is that the function is not differentiable, since the eigenvalues are not differentiable quantities at points where they coalesce. Furthermore, we can usually expect the solution to be at a nondifferentiable point, since the minimization of $\varphi(x)$ will generally drive several eigenvalues to the same minimum value.

In this paper we outline an algorithm that solves (1.1) with an asymptotic quadratic rate of convergence generically. Furthermore, second derivatives are not always required to obtain the quadratic convergence. In order to keep the paper fairly short we will not give proofs of convergence and we will omit some details of the algorithm, but the main ideas should be very clear. We believe this is the first time a quadratically convergent algorithm, or indeed any practical high-accuracy algorithm, has been described for minimizing $\varphi(x)$. An important feature of the algorithm is the ability to obtain a descent direction from any point that is not optimal, even if this requires splitting eigenvalues that are currently equal. (There are exceptions in degenerate cases.) This is also apparently new.

In these respects the algorithm given here represents a significant improvement on the first-order methods for the same problem described by Polak and Wardi (1982) and Doyle (1982). The present paper is heavily influenced by two works, Fletcher (1985) and Friedland, Nocedal, and Overton (1987), to which full acknowledgment is given. Personal communication with Doyle was also very helpful. Another important early reference is Cullum, Donath, and Wolfe (1975), who give a first-order method for a related problem. Undoubtedly a variant of the algorithm given here could be derived for that problem. Finally, we should not overlook the related structural engineering literature (see Olhoff and Taylor (1983, p. 1146) for a useful survey).

2. Connections with the work of Fletcher and Friedland, Nocedal, and Overton. The problem (1.1) may be rewritten as the nondifferentiable constrained optimization problem

$$\min_{\omega \in \mathbb{R}, x \in \mathbb{R}^m} \omega \tag{2.1}$$

$$\text{s.t.} \quad -\omega \leqq \lambda_i(A(x)) \leqq \omega, \qquad i = 1, \cdots, n, \tag{2.2}$$

or equivalently

$$\min_{\omega \in \mathbb{R}, x \in \mathbb{R}^m} \omega \tag{2.3}$$

$$\text{s.t.} \quad \omega I - A(x) \geqq 0, \tag{2.4}$$

$$\omega I + A(x) \geqq 0 \tag{2.5}$$

where "$\geqq$" in (2.4), (2.5) indicates a matrix positive semidefinite constraint. The second formulation immediately suggests that the work of Fletcher (1985) is applicable. Fletcher gives a quadratically convergent algorithm to solve

$$\max \sum_{i=1}^{m} x_k \tag{2.6}$$

$$\text{s.t.} \quad A_0 - \text{Diag}\,(x) \geqq 0, \qquad x \geqq 0 \tag{2.7}$$

and many of the components of his algorithm are therefore applicable to solving (2.3)–(2.5). However, the algorithm is not directly applicable and there are several reasons why it is possible to improve on Fletcher's method in this case. One reason is that Fletcher's method solves a sequence of subproblems, each defined by a guess of the nullity of $A_0 + \text{Diag}\,(x)$, until the correct nullity is identified. Such a strategy cannot easily be extended to the case of two (or more) semidefinite constraints. One goal of our algorithm is to be able to adjust multiplicity estimates while always obtaining a reduction of $\varphi(x)$ at each iteration. We are able to do this by computing an eigenvalue-eigenvector factorization of $A(x)$ at each iteration. By contrast, Fletcher's method uses a block Choleski factorization of $A_0 + \text{Diag}\,(x)$, together with an exact penalty function to impose (2.7).

Also, because of the special form of (2.6), (2.7), Fletcher's method does not require a technique for splitting eigenvalues. In other words, given a matrix $A_0 + \text{Diag}\,(x)$, satisfying (2.7), with nullity t, it cannot be advantageous, in the sense of increasing (2.6), to reduce the multiplicity t. On the other hand it may be necessary to split a multiple eigenvalue in our case, and the ability to recognize this situation and obtain an appropriate descent direction is an important part of our algorithm.

Because we use an eigenvalue factorization of the matrix $A(x)$ at each iterate x, our method has much in common with the methods described by Friedland, Nocedal, and Overton (1987). In the latter paper several quadratically convergent methods are given to solve

$$\lambda_i(A(x)) = \omega, \qquad i = 1, \cdots, t, \tag{2.8}$$

$$\lambda_i(A(x)) = \mu_i, \qquad i = t+1, \cdots, \hat{t} \tag{2.9}$$

where $(\omega, \{\mu_i\})$ are given distinct values and t, $\hat{t}$ (and m, the number of variables) are appropriately chosen. One of the contributions of that paper was to explain that the condition (2.8), although apparently only t conditions, actually generically imposes $t(t+1)/2$ linearly independent constraints on the parameter space, and that effective numerical methods must be based on this consideration. The present paper may be viewed as generalizing the methods of Friedland, Nocedal, and Overton to solve

$$\min_{\omega \in \mathbb{R}, x \in \mathbb{R}^m} \omega \tag{2.10}$$

$$\text{s.t.} \quad \lambda_i(A(x)) = \omega, \qquad i = 1, \cdots, t, \tag{2.11}$$

$$\lambda_i(A(x)) = -\omega, \qquad i = n-s+1, \cdots, n \tag{2.12}$$

where, as a product of the minimization process, it is established that $\omega = \max(\lambda_1, -\lambda_n)$ with

$$\omega = \lambda_1 = \cdots = \lambda_t > \lambda_{t+1} \geqq \cdots \geqq \lambda_{n-s} > \lambda_{n-s+1} = \cdots = \lambda_n = -\omega. \tag{2.13}$$

We shall subsequently refer to t and s as the upper and lower (eigenvalue) multiplicities of $A(x)$. Note that it is possible that either t or s is zero. The following notation will be useful subsequently: let $\{q_1(x), \cdots, q_n(x)\}$ be any orthonormal set of eigenvectors of $A(x)$ corresponding to $\{\lambda_i\}$, and let $Q_1 = [q_1, \cdots, q_t]$, $Q_2 = [q_{n-s+1}, \cdots, q_n]$.

3. Optimality conditions. As Fletcher points out, it is convenient to initially consider the variable space to be the set of all $n \times n$ symmetric matrices $\{A\}$ and to consider the positive semidefinite cone

$$K = \{A | A \geqq 0\}. \tag{3.1}$$

Define an inner product on the set of symmetric matrices by

$$A:B = \operatorname{tr} AB = \sum_{i,j} a_{ij} b_{ij}. \tag{3.2}$$

The normal cone (Rockafellar (1970)) is defined by

$$\partial K(A') = \{B | A':B = \sup_{A \in K} A:B\}. \tag{3.3}$$

Fletcher shows that a very useful expression for ∂K is

$$\partial K(A') = \{B | B = -ZUZ^T, U = U^T, U \geqq 0\} \tag{3.4}$$

where the columns of Z span the null space of A'.

Now consider the restricted variable spaces

$$\hat{K}_1 = \{(\omega, x) | \omega I - A(x) \geqq 0; \omega \in \mathbb{R}; x \in \mathbb{R}^m\}, \tag{3.5}$$

$$\hat{K}_2 = \{(\omega, x) | \omega I + A(x) \geqq 0; \omega \in \mathbb{R}; x \in \mathbb{R}^m\}. \tag{3.6}$$

By definition,

$$\partial \hat{K}_1(\omega', x') = \{(\delta, d) | (\omega', x')^T(\delta, d) = \sup_{(\omega, x) \in \hat{K}_1} (\omega, x)^T(\delta, d)\}. \tag{3.7}$$

THEOREM 3.1.

$$\partial \hat{K}_1(\omega', x') = \{(\delta, d) | \delta = B:I;\ d_k = -B:A_k,\ k = 1, \cdots, m,\ B \in \partial K(\omega' I - A(x'))\}. \tag{3.8}$$

Proof. The proof is omitted because it is almost identical to the proof of Fletcher's Theorem 4.1. Fletcher's proof essentially covers the special case (1.3). One important point worth mentioning is that Fletcher's construction of a feasible arc may require an augmenting term $\alpha\varepsilon^2 I$ in the arc parameterization; in our case this may be absorbed by the ωI term in $\omega I - A(x)$. □

We can now state the optimality condition for x to solve (1.1).

THEOREM 3.2. *A necessary and sufficient condition for x to solve* (1.1) *is that there exist matrices U and V of dimension $t \times t$ and $s \times s$, respectively, with $U = U^T \geqq 0$, $V = V^T \geqq 0$, such that*

$$\operatorname{tr} U + \operatorname{tr} V = 1, \tag{3.9}$$

$$(Q_1^T A_k Q_1):U - (Q_2^T A_k Q_2):V = 0, \qquad k = 1, \cdots, m. \tag{3.10}$$

Here t, s, Q_1, Q_2 are defined by (2.13) and the following remarks.

Proof. Because of the equivalence of (1.1) with the convex problem (2.3)–(2.5), the necessary and sufficient condition for optimality is

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + g_1 + g_2 = 0$$

where $g_1 \in \partial \hat{K}_1$ and $g_2 \in \partial \hat{K}_2$ (Rockafellar (1981, Chap. 5)). By Theorem 3.1 we therefore require

$$1 + \operatorname{tr} B_1 + \operatorname{tr} B_2 = 0,$$

$$-B_1 : A_k + B_2 : A_k = 0, \qquad k = 1, \cdots, m$$

where $B_1 \in \partial K(\omega I - A(x))$, $B_2 \in \partial K(\omega I + A(x))$, $\omega = \max(\lambda_1(A(x)), -\lambda_n(A(x)))$. By (3.4) we have

$$B_1 = -Q_1 U Q_1^T, \qquad B_2 = -Q_2 V Q_2^T$$

for some $U \geqq 0$ and $V \geqq 0$, since Q_1 is a basis for the null space of $\omega I - A(x)$ and Q_2 for $\omega I + A(x)$. Now as Fletcher points out, $U : (Z^T A Z) = A : (Z U Z^T)$ for any $A \in \mathbb{R}^{n \times n}$, $U \in \mathbb{R}^{t \times t}$, and $Z \in \mathbb{R}^{n \times t}$. (A proof is as follows: $U : Z^T A Z = \operatorname{tr} U Z^T A Z = \operatorname{tr} Z(U Z^T A) = (Z U Z^T) : A$, where the middle equality holds because $\operatorname{tr} ZP = \operatorname{tr} PZ$, where $Z \in \mathbb{R}^{n \times t}$, $P \in \mathbb{R}^{t \times n}$.) The theorem is therefore proved. □

The matrices U and V in (3.9) and (3.10) play the role of Lagrange multipliers, as will become clear in the next section. Because the optimality conditions $U \geqq 0$, $V \geqq 0$ are conditions on the matrices as a whole, rather than componentwise conditions, we call U and V Lagrange matrices (cf. "Lagrange vectors" in Overton (1983)).

4. An algorithm based on successive quadratic programming. As explained by Friedland, Nocedal, and Overton, a quadratically convergent method for solving the nondifferentiable system (2.8), (2.9) may be obtained by applying a variant of Newton's method to the nonlinear but essentially differentiable system

$$Q_1(x)^T A(x) Q_1(x) = \omega I_t, \tag{4.1}$$

$$q_i(x)^T A(x) q_i(x) = \mu_i, \qquad i = t+1, \cdots, \hat{t} \tag{4.2}$$

where the columns of $Q_1(x)$ are an orthonormal set of eigenvectors of $A(x)$ corresponding to ω. Here I_t denotes the identity matrix of order t. Let x^* satisfy (4.1), (4.2). Note that (4.1) is independent of the choice of basis for $Q_1(x^*)$. Also note that for points in a neighbourhood of x^*, $A(x)$ will generally have distinct eigenvalues (with small separation) and hence $Q_1(x)$, the matrix of eigenvectors corresponding to the multiple eigenvalue at x^*, will be a well-defined but ill-conditioned function of x which does not converge as $x \to x^*$. This does not cause any difficulties for the Newton method (see Friedland, Nocedal, and Overton (1987) for details). In order to obtain quadratic convergence we need the number of equations, $t(t+1)/2 + (\hat{t} - t)$, to equal the number of variables (together with a nonsingularity condition). When we differentiate (4.1), (4.2), we find that the appropriate system of equations to solve at each step of the Newton method is

$$Q_1(x)^T A(x+d) Q_1(x) = \omega I_t,$$

$$q_i(x)^T A(x+d) q_i(x) = \mu_i, \qquad i = t+1, \cdots, \hat{t}$$

where x is the current iterate and $x + d$ becomes the new iterate. (Although this may look counterintuitive, note that the left-hand side of (4.2) is simply $\lambda_i(x)$, and hence the latter equation is consistent with the well-known fact that the derivative of $\lambda_i(x)$ (with respect to x_k) is $q_i(x)^T A_k q_i(x)$. Again, see Friedland, Nocedal, and Overton (1987) for details.) Since $A(x+d)$ is affine, these equations form a linear system in d. Once $x + d$ is obtained, an eigenvalue-eigenvector factorization of A at the new point is required to be able to start the next iteration.

Now consider generalizing this method to solve (2.10)–(2.12), where we assume for the moment that t and s are known. We see that the Newton method should be applied to the nonlinear problem

$$\min_{\omega \in \mathbb{R}, x \in \mathbb{R}^m} \omega \tag{4.3}$$

$$\text{s.t.} \quad \omega I_t - Q_1(x)^T A(x) Q_1(x) = 0, \tag{4.4}$$

$$\omega I_s + Q_2(x)^T A(x) Q_2(x) = 0. \tag{4.5}$$

The appropriate subproblem to solve at each step of the Newton method is the quadratic program (QP)

$$\min_{\omega \in \mathbb{R}, d \in \mathbb{R}^m} \omega + \tfrac{1}{2} d^T W d \tag{4.6}$$

$$\text{s.t.} \quad \omega I_t - Q_1(x)^T A(x+d) Q_1(x) = 0, \tag{4.7}$$

$$\omega I_s + Q_2(x)^T A(x+d) Q_2(x) = 0 \tag{4.8}$$

where W is a matrix to be specified shortly.

Now define a Lagrangian function for (4.3)–(4.5) by

$$L(\omega, x, U, V) = \omega - U:(\omega I_t - Q_1(x)^T A(x) Q_1(x)) - V:(\omega I_s + Q_2(x)^T A(x) Q_2(x)) \tag{4.9}$$

where $U = U^T$, $V = V^T$. Since (4.7)–(4.8) represent a linearization of (4.4)–(4.5), we see that the first-order necessary condition for x to solve (4.3)–(4.5), namely $\nabla_{\omega,x} L = 0$, is that there exist symmetric matrices U and V such that (3.9)–(3.10) hold—the same optimality condition given in the previous section. (Similarly, if a sequence of QPs (4.6)–(4.8) has been solved, converging to a solution of (4.3)–(4.5) and hence with d converging to zero, the optimality condition of the limiting QP is that there exist U and V such that (3.9)–(3.10) hold.) The equivalence with the optimality conditions (3.9)–(3.10) is very important, since it means that the Lagrange matrices required to check the optimality conditions (3.9), (3.10) may be obtained by solving (4.3)–(4.5), or more specifically, by solving a sequence of QPs (4.6)–(4.8). This observation is the same as the one emphasized by Fletcher and is the essential justification for an algorithm based on successive quadratic programming (SQP). The key point is that (4.3)–(4.5) is much more tractable than the original problem. A related point to note is that U and V are not required to be positive semidefinite for an optimal solution to (4.3)–(4.5), since the constraints are equalities. If U or V is indefinite, this is an indication that t or s is too large and that it is necessary to split a multiple eigenvalue, as will be explained in § 5.

The number of constraints in (4.4), (4.5) is

$$\frac{t(t+1)}{2} + \frac{s(s+1)}{2}. \tag{4.10}$$

If this quantity is equal to $m + 1$ (the number of variables in (4.3)), then, generically, the constraints themselves are enough to define a unique solution to (4.3)–(4.5) and the SQP method will have local quadratic convergence regardless of the value of W. In this case the solution of (1.1) is "strongly unique." If (4.10) is greater than $m + 1$, then, except in degenerate cases, (4.3)–(4.5) will be infeasible. In general we expect (4.10) to be less than or equal to $m + 1$, but we cannot expect equality—for example, if $m = 4$, equality is not possible. If (4.10) is less than $m + 1$ the proper choice of the matrix W is necessary for the SQP method to converge quadratically. It is clear that W should be set to the

Hessian, with respect to x, of the Lagrangian function (4.9). It can be shown that this matrix is given by

$$W_{jk} = U{:}G_1^{j,k} - V{:}G_2^{j,k} \tag{4.11}$$

where

$$G_l^{j,k} = 2Q_l(x)^T A_k \bar{Q}_l(x)(\omega J_l - \bar{\Lambda}_l(x))^{-1} \bar{Q}_l(x)^T A_j Q_l(x), \qquad l = 1, 2, \tag{4.12}$$

and where the columns of $\bar{Q}_l(x)$ consist of all eigenvectors in $\{q_1, \cdots, q_n\}$ except those in $Q_l(x)$ ($l = 1, 2$), $\bar{\Lambda}_l(x)$ is a diagonal matrix whose entries consist of all eigenvalues in $\{\lambda_1, \cdots, \lambda_n\}$ except those corresponding to $Q_l(x)$, and $J_1 = I_{n-t}$, $J_2 = -I_{n-s}$.

Some caution is required in the choice of U and V in (4.11). Since the values of U and V at the minimum of (4.3)–(4.5) are not known, it is necessary to use Lagrange matrix estimates. The obvious choice is to use the values obtained from the previous QP. Unfortunately these are useless, because the eigenvector bases $Q_1(x)$ and $Q_2(x)$ at the current point will generally have no relation to those at the previous point (although the range spaces of $Q_1(x)$ and $Q_2(x)$ will converge as x converges to a minimizing point). Therefore, after $Q_1(x)$, $Q_2(x)$ have been computed, but before solving the QP, it is necessary to obtain first-order Lagrange matrix estimates by minimizing the 2-norm of the residual of (3.9), (3.10). This does not require a significant amount of extra work since the QR factorization of the relevant coefficient matrix is needed anyway to solve the QP. (See Murray and Overton (1980) for some comments on Lagrange multiplier estimates for minimax problems and Nocedal and Overton (1985) for comments on first- and second-order Lagrange multiplier estimates.) Alternatively, we could use the Cayley transform method ("Method III") of Friedland, Nocedal, and Overton, which updates estimates of the eigenvectors without recomputing them in such a way that even the eigenvector estimates corresponding to multiple eigenvalues converge. This technique does not impede quadratic convergence. It would be essential if we wanted to use a quasi-Newton method to approximate the matrix W without computing (4.11), which might be necessary for large problems. Note again, however, that W is not needed at all if (4.10) equals $m + 1$, which may quite often be the case.

We now turn to the important question of how the upper and lower multiplicities t and s are to be determined. These can be effectively obtained dynamically. Suppose that $t = 1$, $s = 0$ initially. If the QP (4.6)–(4.8) were now to be solved, the solution would very likely reduce the initially largest eigenvalue far below the others. It is therefore sensible to incorporate into the QP inequality constraints on the other eigenvalues, namely

$$-\omega \leqq q_i^T(x) A(x + d) q_i(x) \leqq \omega, \qquad t + 1 \leqq i \leqq n - s. \tag{4.13}$$

We may now obtain updated estimates of t and s by seeing which constraints are active at the solution of the QP (4.6)–(4.8), (4.13). A reasonable strategy is to increase t by the number of constraints which are at their upper bound and to increase s by the number at their lower bound. However, some caution should be used, since if t and s become too large, (4.3)–(4.5) will become infeasible. We therefore also keep more conservative estimates $\bar{t}$ and $\bar{s}$ which are defined at the beginning of each iteration by

$$\omega - \lambda_i(x) \leqq \text{TOL}, \qquad i = 1, \cdots, \bar{t}, \tag{4.14}$$

$$\omega + \lambda_i(x) \leqq \text{TOL}, \qquad i = n - \bar{s} + 1, \cdots, n \tag{4.15}$$

assuming that $\bar{t}(\bar{t} + 1)/2 + \bar{s}(\bar{s} + 1)/2 \leqq m + 1$, where $\omega = \varphi(x)$ and TOL is a reasonably small number, e.g., 10^{-2}. If necessary t and s are reset to these more conservative values, as will be explained shortly. However, if t and s are always set to $\bar{t}$ and $\bar{s}$ instead of making use of the active constraint information from the solution of the previous QP, the al-

gorithm, though reliable, converges much more slowly, since quadratic convergence cannot take place until the eigenvalue separation is reduced to TOL (unless the solution has distinct eigenvalues). (A possible alternative approach to accelerating the selection of the correct multiplicity estimates would be to use a special line search as is done in Overton (1983).)

At each iteration we insist that a reduction in ω is obtained. Even if t and s have the correct values defined by (2.13), there is no guarantee that the solution d of (4.6)–(4.8), (4.13) will give $\varphi(x + d) < \varphi(x)$. Following Fletcher, we therefore use a "trust region" strategy, incorporating into the QP bound constraints

$$|d_k| \leqq \rho, \qquad k = 1, \cdots, m \tag{4.16}$$

where ρ is dynamically adjusted. It is clear that if ρ and TOL are sufficiently small, then the solution d of the QP (4.6)–(4.8), (4.13), (4.16), with $t = \bar{t}$, $s = \bar{s}$, will give $\varphi(x + d) < \varphi(x)$ unless $d = 0$.

If TOL $= 0$, $t = \bar{t}$, $s = \bar{s}$ and the solution d of the QP is zero, the point x is a minimum of (4.3)–(4.5). It therefore also solves (1.1) if the Lagrange matrices U and V are positive semidefinite. If U or V is indefinite then it is both necessary and feasible to split a multiple eigenvalue to make further progress, as will be explained in § 5.

We conclude this section with a summary of the algorithm. It requires initial values for TOL and ρ and a convergence tolerance ε.

ALGORITHM.

0. Given x, evaluate $\{\lambda_i(x)\}$, $\{q_i(x)\}$. Define $\bar{t}$, $\bar{s}$ by (4.14), (4.15). Set $t = \bar{t}$, $s = \bar{s}$.
1. Solve the QP (4.6)–(4.8), (4.13), (4.16), using first-order Lagrange matrix estimates to define W. If the QP is infeasible, go to Step 2.2. If $\|d\| \leqq \varepsilon$, go to Step 3.
2. Evaluate $\{\lambda_i(x + d)\}$. If $\varphi(x + d) < \varphi(x)$, then
   2.1 Increase t and s, respectively, by the number of upper and lower bounds which are active in (4.13), provided the new values give (4.10) less than or equal to $m + 1$. Set x to $x + d$, evaluate $\{q_i(x)\}$, and define $\bar{t}$, $\bar{s}$ by (4.14), (4.15). Double ρ, and go to Step 1.

else

   2.2 Reset t, s to $\bar{t}$, $\bar{s}$. Divide ρ by two and go to Step 1.

3. If $U \geqq 0$ and $V \geqq 0$ then
   3.1 STOP $-$ x is optimal.

else

   3.2 Split a multiple eigenvalue and obtain reduction as described in the next section. Adjust $\bar{t}$, $\bar{s}$, t, s accordingly and go to Step 1.

This algorithm has worked well in practice (see the results in § 6). Clearly it can be defeated; in particular, if TOL is not small enough, the QP may be infeasible, and at present there is no facility for reducing TOL. However, it seems likely that it will form the basis of a more elaborate algorithm for which global convergence can be guaranteed. Because $\varphi(x)$ is convex, obtaining a globally convergent algorithm is not difficult; what is wanted is a globally convergent algorithm for which final quadratic convergence is guaranteed (given nonsingularity assumptions).

5. Splitting multiple eigenvalues. Consider a simple example. Let $m = n = 2$, with

$$A_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & \kappa \\ \kappa & 4 \end{bmatrix} \tag{5.1}$$

for some value κ. Since $A(x) = A_0 + x_1A_1 + x_2A_2$, the only point where $A(x)$ has multiple eigenvalues is $x = (0, 0)^T$, which is therefore a solution of (4.3)–(4.5) with $t = 2$, $s = 0$. If κ is large enough, clearly $x = (0, 0)^T$ is a minimum of $\varphi(x)$, since $x_1A_1 + x_2A_2$ is indefinite for any nonzero x. On the other hand, if κ is small enough, A_2 is positive definite, and $d = (0, -1)^T$ is a descent direction from $x = (0, 0)^T$. It is therefore essential to be able to distinguish between these situations and to find a descent direction if one exists. It appears that an inability to do this has been one of the major deficiencies of algorithms previously developed for (1.1) (Doyle (1986)).

In order to check optimality, we introduce the Lagrange matrix U (V is empty since $s = 0$). The system (3.9), (3.10) is

$$\begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 0 \\ -1 & -4 & -\kappa \end{bmatrix} \begin{bmatrix} U_{11} \\ U_{22} \\ 2U_{12} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

where we arbitrarily choose $Q_1 = I$. The solution is

$$U = \begin{bmatrix} \frac{1}{2} & \frac{-5}{4\kappa} \\ \frac{-5}{4\kappa} & \frac{1}{2} \end{bmatrix}. \tag{5.2}$$

The optimality condition is $U \geqq 0$, i.e.,

$$|\kappa| \geqq \frac{5}{2}.$$

We now show how to obtain a descent direction if $|\kappa| < \frac{5}{2}$. The solution is to solve

$$\delta I - \sum_{k=1}^{2} d_k A_k = -\mu u u^T$$

where μ is the negative eigenvalue of U and u is the corresponding eigenvector. This gives, in the case of $\kappa = 2.25$,

$$\begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & -4 \\ 0 & 0 & -\kappa \end{bmatrix} \begin{bmatrix} \delta \\ d_1 \\ d_2 \end{bmatrix} = 2.78 \times 10^{-2} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

i.e., $\delta = -3.09 \times 10^{-3}$, $d = (-1.85 \times 10^{-2}, -1.23 \times 10^{-2})^T$. Now $\lambda(A(x + d)) = (0.941, 0.997)^T$ so that $\varphi(x + d) < \varphi(x)$ as required. Note that $d = (0, -1)^T$ is not a descent direction from $x = 0$ in this case.

More generally, we have the following.

THEOREM 5.1. *Let t and s be defined by* (2.13). *Assume x is a minimum of* (4.3)–(4.5), *so that* (3.9), (3.10) *hold for some symmetric matrices $U \in \mathbb{R}^{t \times t}$ and $V \in \mathbb{R}^{s \times s}$. Suppose that U is indefinite with a negative eigenvalue μ and corresponding eigenvector u. Then if (δ, d) solves*

$$\delta I_t - \sum_{k=1}^{m} d_k Q_1^T A_k Q_1 = -\mu u u^T, \tag{5.3}$$

$$\delta I_s + \sum_{k=1}^{m} d_k Q_2^T A_k Q_2 = 0, \tag{5.4}$$

we have that d is a descent direction for $\varphi(x)$. *Furthermore, to first order the multiplicity t is reduced by exactly one along d, and the new set of eigenvectors for* $\lambda_i = \omega$ *can be taken, to first order, as*

$$\tilde{Q}_1 = Q_1 \bar{U}$$

where the columns of $\bar{U}$ *are the eigenvectors of U, excluding u.*

Remark. Equations (5.3)–(5.4) are generically solvable if (4.10) is less than or equal to $m + 1$. Other cases are degenerate situations for which obtaining a descent direction is more difficult.

Proof. Taking an inner product of U with (5.3) and V with (5.4) and adding them together we obtain

$$\delta(\operatorname{tr} U + \operatorname{tr} V) + \sum_{k=1}^{m} d_k(-U:Q_1^T A_k Q_1 + V:Q_2^T A_k Q_2) = -\mu^2.$$

It therefore follows from (3.9), (3.10) that

$$\delta = -\mu^2. \tag{5.5}$$

Furthermore, for the same reason that (4.7) is a valid linearization of (4.4), (5.3), and (5.4) show that the constraints (2.4), (2.5) hold to first order along the direction $x + \alpha d$, $\alpha \geqq 0$ (since the right-hand sides of (5.3), (5.4) are positive semidefinite). It follows from (5.5) that d is a descent direction. Finally, the last statement is justified by multiplying (5.3) by $(u, \bar{U})^T$ on the left and $(u, \bar{U})$ on the right, obtaining

$$\delta I_t - \sum_{k=1}^{m} d_k(u, \bar{U})^T Q_1^T A_k Q_1 (u, \bar{U}) = \begin{bmatrix} -\mu & \\ & 0 \end{bmatrix}. \qquad \square$$

In other words, all eigenvalues but one are reduced by μ^2 (to first order) while the other eigenvalue is reduced by $\mu^2 - \mu$.

More generally still, if U has more than one negative eigenvalue (or U and V both have negative eigenvalues), we can reduce t by more than one (or reduce both t and s) by replacing the right-hand side of (5.3) (and (5.4)) by a sum of outer products corresponding to the negative eigenvalues. This has an obvious analogy in nonlinear programming, where if several Lagrange multipliers are negative at a stationary point we can move off just a single constraint (as does the simplex method for linear programming) or move off several constraints at once. Also, in nonlinear programming we may move off a constraint before minimizing on the corresponding manifold if the appropriate Lagrange multiplier estimate is negative. Similarly, we should be able to use Lagrange matrix estimates to avoid minimizing on the manifold defined by (4.4), (4.5).

6. Numerical examples. The algorithm has been implemented in Fortran and run on a Pyramid Unix system at Australian National University. Double precision arithmetic (about 15 decimal digits of accuracy) was used. The eigensystems of $A(x)$ were computed using EISPACK (Smith et al. (1967)). The QPs were solved using the Stanford package QPSOL (Gill et al. (1984)).

We give three examples that illustrate the effectiveness of the method. The parameters ε and TOL were given the values 10^{-7} and 10^{-2}, respectively, and the initial trust region radius ρ was set to 1. The tables shown below have the following meaning. There is one row in the table for each time a reduction in $\varphi(x + d)$ is obtained, i.e., Step 2.1 is executed. The values $\bar{t}$, $\bar{s}$, t, and s are those holding at the beginning of the iteration, i.e., following the previous execution of Step 0 or 2.1. The quantity #QPs is the number of QPs that

had to be solved before obtaining a reduction, i.e., the number of times Step 1 was executed. Step 3.2 was not executed in any of these examples.

Example 1. This is defined by (5.1) with $\kappa = 3$.

Initial $x = (1.0, 2.0)^T$ with $\varphi(x) = 12.32$.

Iteration	$\bar{t}$	$\bar{s}$	t	s	#QPs	$\varphi(x+d)$
1	1	0	1	0	1	6.541381
2	1	0	1	0	1	4.817767
3	1	0	2	0	1	1.000000

Final $x = (0.0, 2.0 \times 10^{-15})^T$ with $\lambda(x) = (1.0, 1.0)^T$ and

$$U = \begin{bmatrix} 0.8716 & -0.1884 \\ -0.1884 & 0.1284 \end{bmatrix}.$$

Comments. Once the correct multiplicities are identified this particular problem is solved in one step. The reason that U is different from (5.2) is that EISPACK chose a basis $Q_1 \neq I$. Of course this does not affect the optimality condition.

Example 2. $n = 3$, $m = 3$,

$$A_0 = \begin{bmatrix} 0 & 1.0 & 1.1 \\ 1.0 & 0 & 1.2 \\ 1.1 & 1.2 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 1 \end{bmatrix}.$$

Initial $x = (1.0, 0.9, 0.8)^T$ with $\varphi(x) = 7.605$.

Iteration	$\bar{t}$	$\bar{s}$	t	s	#QPs	$\varphi(x+d)$
1	1	0	1	0	1	1.616283
2	1	0	1	0	1	1.464941
3	0	1	1	2	1	1.145090
4	1	1	1	2	1	1.102385
5	1	2	1	2	1	1.101521
6	1	2	1	2	1	1.101520

Final $x = (-0.1163679, -0.2497934, -1.845989)^T$ with

$$\lambda(x) = (1.101520, -1.101520, -1.101520)^T$$

and

$$U = [6.95 \times 10^{-4}], \qquad V = \begin{bmatrix} 0.4861 & 0.0229 \\ 0.0229 & 0.5132 \end{bmatrix}.$$

Comments. Note that U is only barely positive definite, so that a small perturbation to the problem would give an optimal point with $\lambda_1 < \omega$. As in Example 1, (4.10) equals $m + 1$ at the solution, so W is not needed for quadratic convergence, although it may help to identify t and s during the early iterations. Note also that following the first iteration where the correct multiplicities were used to define the QP, the solution is correct to two figures.

Example 3. $n = 10$, $m = 10$, $A_k = e_k e_k^T$, $k = 1, \cdots, 10$, and

$$A_0 = \begin{bmatrix} 0 & & & & & & & & & \\ 1.1 & 0 & & & & & & & & \\ 1 & 2.1 & 0 & & & & & \text{(transpose)} & & \\ 1 & 2 & 3.1 & 0 & & & & & & \\ 1 & 2 & 3 & 4.1 & 0 & & & & & \\ 1 & 2 & 3 & 4 & 5.1 & 0 & & & & \\ 1 & 2 & 3 & 4 & 5 & 6.1 & 0 & & & \\ 1 & 2 & 3 & 4 & 5 & 6 & 7.1 & 0 & & \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8.1 & 0 & \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9.1 & 0 \end{bmatrix}.$$

Initial $x = (1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1)^T$ with $\varphi(x) = 38.09$.

Iteration	$\bar{t}$	$\bar{s}$	t	s	#QPs	$\varphi(x + d)$
1	1	0	1	0	1	37.08646
2	1	0	1	0	1	35.08646
3	1	0	1	0	1	31.08646
4	1	0	1	0	1	23.30168
5	1	1	1	1	1	23.06948
6	0	1	1	1	7	22.57218
7	1	1	1	1	3	22.55570
8	0	1	1	2	2	22.43732
9	0	1	1	3	3	22.39628
10	0	1	1	3	2	22.37459
11	1	2	1	2	1	22.37020
12	1	2	1	2	1	22.36642
13	1	2	1	2	1	22.36613
14	1	2	1	2	1	22.36612

Final $x = (-21.25583, -20.58868, -19.24580, -18.60455, -17.22383, -16.63475, -15.18517, -14.74159, -13.05307, -13.46085)^T$ with

$$\lambda(x) = (22.36612, -17.32323, -20.48036, -21.34962, -21.69938, -22.17358, -22.26831, -22.33351, -22.36612, -22.36612)^T$$

and

$$U = [0.5], \quad V = \begin{bmatrix} 0.3445 & -5.017 \times 10^{-3} \\ -5.017 \times 10^{-3} & 0.1555 \end{bmatrix}.$$

Comments. This problem is quite difficult to solve, since at the solution the interior eigenvalues are nearly equal to λ_n. Indeed, if a larger value of TOL had been used, the QP probably would have become infeasible making it necessary to reduce TOL. During the first few iterations, larger improvements were inhibited by the trust region radius, which was successively doubled. At iteration 5 the QP solution indicated that t, s should be set to 1, 9, but since this would have made (4.10) greater than $m + 1$, t and s were not increased. As a result, seven QPs were required during iteration 6 until the trust radius was small enough to make progress. Eventually quadratic convergence was obtained once the correct multiplicities were identified. In this case the second derivative matrix W was essential for quadratic convergence.

In general we would not expect Step 3.2 to be required. The reason for this is that when t or s is increased to a value $\geqq 2$, the iterate x is essentially moving onto a manifold which has dimension at least two lower than the current constraining manifold. This is unlikely to happen by accident, but only likely to occur in the course of making progress towards optimality. However, the ability to split multiple eigenvalues is still important in case it is needed because of starting at an unfortunate point or in the course of solving ill-conditioned problems.

7. Final comments. A number of problems in addition to (1.1) may be solved by related techniques. Clearly it is trivial to extend the algorithm given here to solve

$$\min_x \max_{1 \leqq l \leqq p} \max_{1 \leqq i \leqq n} |\lambda_i(A^{(l)}(x))|,$$

where $A^{(1)}(x), \cdots, A^{(p)}(x)$ are each affine matrix-valued functions, by simply introducing additional constraints to the QP and corresponding Lagrange matrices. The algorithm could also be extended to solve more general optimization problems involving constraints on eigenvalues of various matrix functions. It would be necessary to introduce a penalty function to measure progress towards the solution. Constraints on interior eigenvalues could also be included (although these would not be convex).

Finally, it is possible to extend the algorithm to handle nonlinear matrix functions $A(x)$, although the resulting optimization problem is no longer necessarily convex. The necessary changes are mainly to replace A_k by $\partial A(x)/\partial x_k$ in the derivative formulas, and to be aware of the need to verify second-order optimality conditions.

REFERENCES

J. CULLUM, W. E. DONATH, AND P. WOLFE (1975), *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Programming Study, 3, pp. 35–55.

J. DOYLE (1982), *Analysis of feedback systems with structured uncertainties*, IEEE Proc., 129, pp. 242–250.

——— (1986), private communication.

R. FLETCHER (1985), *Semi-definite matrix constraints in optimization*, SIAM J. Control Optim., 23, pp. 493–513.

S. FRIEDLAND, J. NOCEDAL, AND M. L. OVERTON (1987), *The formulation and analysis of numerical methods for inverse eigenvalue problems*, SIAM J. Numer. Anal., 24, pp. 634–667.

P. E. GILL, W. MURRAY, M. A. SAUNDERS, AND M. H. WRIGHT (1984), *User's Guide to QPSOL: a Fortran package for quadratic programming*, Systems Optimization Laboratory Report, Stanford University, Stanford, CA.

W. MURRAY AND M. L. OVERTON (1980), *A projected Lagrangian algorithm for nonlinear minimax optimization*, SIAM J. Sci. Statist. Comput., 1, pp. 345–370.

J. NOCEDAL AND M. L. OVERTON (1985), *Projected Hessian updating algorithms for nonlinearly constrained optimization*, SIAM J. Numer. Anal., 22, pp. 821–850.

N. OLHOFF AND J. E. TAYLOR (1983), *On structural optimization*, J. Appl. Mech., 50, pp. 1138–1151.

M. L. OVERTON (1983), *A quadratically convergent method for minimizing a sum of Euclidean norms*, Math. Programming, 27, pp. 34–63.

E. POLAK AND Y. WARDI (1982), *Nondifferentiable optimization algorithm for designing control systems having singular value inequalities*, Automatica, 18, pp. 267–283.

R. T. ROCKAFELLAR (1970), *Convex Analysis*, Princeton University Press, Princeton, NJ.

——— (1981), *The Theory of Subgradients and Its Applications to Problems of Optimization: Convex and Nonconvex Functions*, Research and Education in Mathematics 1, Heldermann-Verlag, Berlin.

B. T. SMITH, J. M. BOYLE, J. J. DONGARRA, B. S. GARBOW, Y. IKEBE, V. C. KLEMA, AND C. B. MOLER, (1967), *Matrix Eigensystem Routines—EISPACK Guide*, Lecture Notes in Computer Science 6, Springer-Verlag, Berlin, New York.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

012

HYPERBOLIC HOUSEHOLDER TRANSFORMS*

CHARLES M. RADER† AND ALLAN O. STEINHARDT†‡

Abstract. A class of transformation matrices, analogous to the Householder matrices, is developed with a nonorthogonal property designed to permit the efficient deletion of data from least-squares problems. These matrices, which we term hyperbolic Householder, are shown to effect deletion, or simultaneous addition and deletion, of data with much less sensitivity to rounding errors than for techniques based on normal equations. When the addition/deletion sets are large, this numerical robustness is obtained at the expense of only a modest increase in computations, and when only a relatively small fraction of the data set is modified, there is a decrease in required computations. Two applications to signal processing problems are considered. First, these transformations are used to obtain a square root algorithm for windowed recursive least-squares filtering. Second, the transformations are employed to implement the rejection of spurious data from the weight vector estimation process in an adaptive array.

Key words. hyperbolic transformations, Householder matrices, *QR* decompositions, Robust Least Squares

AMS(MOS) subject classification. 65F25

1. Introduction. In this paper, we will present a method by which we can solve a succession of least-squares problems, where the data sets of the successive problems have some data in common and other data that is different. Our application area is in the field of adaptive antennas. The solution to an adaptive antenna problem is a vector of weights such that the energy of interference in a certain weighted sum is minimum. We must solve this problem in real-time and then solve it again, and again, with newly observed interference, because the interference is not expected to be stationary. However, we expect the statistics of the interference to change slowly with time, so each time we update the solution of the weight vector we reuse a large fraction of the old observations of interference, bringing in only a few new observations and discarding only a few old observations. This is an example of an application in which we must solve a succession of least-squares problems, each involving the deletion of some old data and the insertion of some new data.

Another type of problem is one in which we solve a certain least-squares problem with a given data set, then discover that some of the data is, in some way, spurious. We would like to then resolve the problem with a smaller data set, some of the data deleted, but making as much use as possible of our earlier work.

The classical method of solving least-squares problems, dating back at least to the time of Gauss, is the method of normal equations. The coefficients of the normal equations, called a correlation matrix, are found by averaging certain products of the raw data, after which the correlation matrix must be inverted. One way to invert the correlation matrix is to first factor it into the matrix product of two triangular matrices which are conjugate transposes of one another. The inverses of triangular matrices are easy to compute. We shall refer to this type of method as the Cholesky/power method. (There

* Received by the editors March 2, 1987; accepted for publication October 1, 1987. This work was sponsored by the Department of the Air Force. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government. This is a condensed version of a paper published in the *IEEE Transactions on Acoustics, Speech, and Signal Processing*. The earlier sections of this paper are largely tutorial and may be omitted by readers who are already familiar with the use of Householder transformation in the solution of least-squares problems. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† MIT Lincoln Laboratory, Lexington, Massachusetts 02173.

‡ Present address, Department of Electrical Engineering, Cornell University, Ithaca, New York 14853.

are many least-squares problems in digital signal processing whose special structure permits more efficient means of solution. For example, when the normal equations are Toeplitz, we may solve them by Levinson recursion. In this paper we do not assume any special structure in the normal equations, but we must alert the reader that when such structure is present, the methods presented in this paper may be relatively much less valuable, although they are still applicable.) There are numerical problems associated with the Cholesky/power method, or with any method which forms the correlation matrix; the elements of the correlation matrix have twice as much dynamic range as the original data. The large dynamic range to be expected in the raw data of an adaptive antenna interference cancellation system is such that it is numerically untenable to work with the dynamic range of the squared data required in the coefficients of the normal equations. Instead of forming the normal equations, we prefer to transform the raw problem data by a series of linear transformations, each of which gives a new data set with the same normal equations as the original data set. Because the normal equations are unchanged by the linear transformations, the solution of the least-squares problem is also unchanged. We ultimately find a data set in the form of a triangular matrix, from which the least-squares solution can be found by simple methods. This triangular matrix is the same as the triangular factor of the correlation matrix, found by the Cholesky/power method, but the numerical problems caused by doubling of dynamic range are avoided.

The allowed linear transformations of raw data may be expressed as the postmultiplication of a raw data matrix by any orthonormal matrix. The Householder matrices are one class of orthonormal matrices. For any given raw data matrix it is easy to construct a Householder matrix which transforms the raw data so that the resulting matrix has zeros in certain positions. Householder transformations are well known in the literature and have been used extensively for the solution of the least-squares problem [1]. The section of this paper devoted to the definition and application of Householder transformations (§ 2) is therefore essentially review. We show, by simulation, that when we apply Householder transformations we can get solutions to our realistic antenna problem using a shorter computer wordlength than we need when we use the conventional approach via the normal equations. In fact, there is nearly a halving in wordlength with the Householder approach. The reduction in wordlength obtained is well worth the associated doubling (or lesser increase) in computations for many applications.

But if we have already invested computational effort in the solution of a least-squares problem involving tens or hundreds of data vectors, we can appreciate any method that allows us to reuse old computations to solve a new problem, when we introduce only a few new data vectors and/or delete only a few old (or unrepresentative) vectors from the data set. Although it is simple to make the changes in a correlation matrix equivalent to incorporating new vectors or deleting old vectors, it is not as simple to make the equivalent changes in the triangular factors described above. Previous workers have described methods of "updating" triangular factors to reflect the incorporation of new data vectors, but have only described "downdating" methods which remove the effect of one vector at a time. It would be convenient to be able to compute a new triangular factor corresponding to simultaneously bringing in many new data vectors and deleting many old data vectors. For this purpose, we introduce a class of matrices which we call hyperbolic Householder matrices.

Consider a composite data matrix which consists of an "old" (triangular) data matrix (derived from an earlier problem) to which is appended a matrix of "new" data and another matrix of "obsolete" data. A hyperbolic Householder matrix describes a linear transformation which, when applied to this composite data matrix, will give us another composite data matrix of the same type, but with more zero-valued elements. Specifically,

the new composite matrix, while containing more zeros, preserves unchanged the sum of the correlation matrix of the initial "old" data matrix and the correlation matrix of the "new" data matrix minus the correlation matrix of the "obsolete" data matrix. After a series of such transformations, we will have a new composite matrix whose transformed "new" and "obsolete" data matrices are all zero, and whose nonzero entries are lower triangular. Therefore, the lower triangular matrix is the desired Cholesky factor for the updated problem. In the third section of the paper, we explain the mathematics of this updating, we count up the computations required to implement it, and we study, via analysis and simulations, its numerical stability. We find that the hyperbolic Householder approach offers a substantial reduction in wordlength (in comparison to updating the normal equations directly). Furthermore, in some important cases, the hyperbolic Householder method offers a reduction in computations as well.

The methods in this paper are all "voltage-domain" methods, in that the doubling of dynamic range associated with the normal equations is avoided. Such methods are frequently referred to as "square root" methods since the triangular factor which we work with is (in some sense) the square root of the original correlation matrix.

Householder matrices are orthonormal[1] matrices. This allows us to put tight bounds on the amount of error introduced into the original problem by the transformation matrices. But the hyperbolic Householder matrices are not orthonormal. A numerical effect analogous to the classical difference-of-large-numbers problem is always a potential hazard when we attempt to account for the deletion of data from a data set. The hyperbolic Householder approach to data deletion does not necessarily avoid this potentially hazardous numerical effect. However, in the course of constructing the hyperbolic Householder matrices, we shall compute intermediate quantities which easily indicate when these effects are present. We show how to compute the ratio of the largest and smallest eigenvalues of a hyperbolic Householder matrix from these intermediate quantities. In turn from these eigenvalues, we develop bounds on the amount of rounding errors introduced into the least-squares solution. These bounds are seldom tight, but simulations indicate that the bounds are useful in predicting when noticeable quantization error arises. We present simulations that demonstrate that significantly lower wordlengths give adequate results when the hyperbolic Householder method is employed in place of updating and downdating the normal equations, even when ill-conditioning arises.

2. Householder transformation matrices and least-squares problems.

2.1. Definition and properties of Householder matrices. Let $\mathbf{B}$ be any complex column vector (with N elements) and let $(\ )^t$ denote conjugate transpose. Then $\mathbf{B}^t\mathbf{B}$ is a real scalar and $\mathbf{B}\mathbf{B}^t$ is a square $(N \times N)$ matrix. Let I be the identity matrix of the same dimensions. Then

$$Q = I - 2\frac{\mathbf{B}\mathbf{B}^t}{\mathbf{B}^t\mathbf{B}} \tag{1}$$

is called a Householder (reflection) matrix. Q is Hermitian and orthonormal. If a Householder matrix is used to multiply a vector or another matrix, its effect on that vector or matrix is called a Householder transformation. When any orthonormal matrix premultiplies a column vector, it leaves the energy in the resulting column vector the same as the energy of the original vector. For any given vector $\mathbf{U}$, we can construct a Householder matrix so that all this energy is compacted into a selected component. We can do this as follows: Let $\mathbf{E}_j$ be the column vector whose jth component (the component into

[1] A complex matrix which satisfies $Q^tQ = I$ is sometimes referred to as unitary rather than orthonormal.

which the energy is to be compressed) is unity and whose other components are all zeros. Then set

$$\mathbf{B} = \mathbf{U} + \sigma \mathbf{E}_j \tag{2}$$

where

$$\sigma = \frac{\pm u_j}{|u_j|} \sqrt{\mathbf{U}^t \mathbf{U}}. \tag{3}$$

Then we can show that

$$Q\mathbf{U} = \mathbf{U} - (\mathbf{U} + \sigma \mathbf{E}_j) = -\sigma \mathbf{E}_j \tag{4}$$

as desired. We will (somewhat arbitrarily) choose the plus sign when forming σ. Equation (3) tells us that the complex number σ has the angle (argument) of u_j and the magnitude of $\mathbf{U}$.

Taking the conjugate transpose of both sides of (4), and noting that Q is Hermitian, we find that

$$\mathbf{U}^t Q = -\sigma^* \mathbf{E}_j^t. \tag{5}$$

We have thus far shown how to construct an orthonormal matrix which, by post-multiplication, compresses all the energy in a particular row vector into its jth entry. We now show how such matrices can be employed in the stable solution of least-squares problems.

2.2. The solution of least-squares problems using Householder transformations. Suppose we are given raw data in the form of an $N \times M$ matrix X, of complex numbers, and are then asked to find an "optimum" N-element vector $\mathbf{W}$. The optimization criterion involves the "outputs" y_n, arranged into a vector $\mathbf{Y}$ with M elements, related to $\mathbf{W}$ and X by

$$\mathbf{Y}^t = \mathbf{W}^t X. \tag{6}$$

The quantity to be minimized is the sum of the squared magnitudes of the components of $\mathbf{Y}$

$$\mathscr{E} = \mathbf{Y}^t \mathbf{Y} = \sum_{n=1}^{M} y_n y_n^* = \mathbf{W}^t X X^t \mathbf{W}. \tag{7}$$

Generally, there will be some other linear constraints on the choice of $\mathbf{W}$; otherwise we would clearly choose $\mathbf{W} = (0, 0, \cdots 0)^t$. These linear constraints will take the form

$$A\mathbf{W} = \mathbf{C} \tag{8}$$

for a given matrix A and a given vector $\mathbf{C}$. But we will concentrate on the special case with only a single constraint—then A is a single row, $\mathbf{V}^t$, and $\mathbf{C}$ becomes a scalar which we can set to 1 with no loss in generality. The straightforward method of solving this minimization problem is to introduce a Lagrange variable ρ and the new function to be minimized is

$$\mathscr{E} = \mathbf{W}^t (X X^t) \mathbf{W} + \rho(\mathbf{V}^t \mathbf{W} - 1). \tag{9}$$

The correlation matrix is XX^t, which we abbreviate by R. It is the only way in which the raw data X enters into the problem. We take derivatives with respect to the components

of $\mathbf{W}$ and with respect to the Lagrange variable ρ and we set them to zero, giving us the set of linear equations

$$R\mathbf{W} + \rho\mathbf{V} = 0, \tag{10}$$

$$\mathbf{V}^t\mathbf{W} = 1. \tag{11}$$

We solve these equations by setting

$$\mathbf{W} = -\rho R^{-1}\mathbf{V} \tag{12}$$

where the unknown ρ can initially be set to -1 and the resulting solution then can be scaled using (11).

The above discussion shows that $\mathbf{W}$ is the solution (to within a scale factor) of the equation

$$(XX^t)\mathbf{W} = \mathbf{V}. \tag{13}$$

If we can find an equivalent set of data, $\hat{X}$, in the sense that its correlation matrix is the same,

$$XX^t = \hat{X}\hat{X}^t = R, \tag{14}$$

then the same solution vector $\mathbf{W}$ would apply; in particular, if $\hat{X}$ has the form of a triangular matrix, it is then fairly easy to solve (13) because triangular matrices $\hat{X}$ and $\hat{X}^t$ are fairly easy to invert. Equation (14) also tells us that it is valid to apply certain linear transformations to the given data X as long as the transformed data, $\hat{X} = XQ$, has the same correlation matrix as the original data X,

$$(XQ)(XQ)^t = XQQ^tX^t = XX^t = R. \tag{15}$$

We see that the allowed transformations are the postmultiplication of the raw data array by any orthonormal matrix. We can build Q as a product of Householder matrices. To begin, let Q_1 be formed via (1) and (2), with $j = 1$ and $\mathbf{U}$ set to the Hermitian transpose of the first row of X. Then $X^1 = XQ_1$ has a first row with only one nonzero entry, which is in the first column.

Note that it is much less expensive to postmultiply X by a Householder matrix Q than it would be to postmultiply it by a general $M \times M$ matrix. Consider the following procedure:

$$\mathbf{G} \leftarrow X\mathbf{B}, \tag{16}$$

$$\mathbf{S} \leftarrow \frac{2\mathbf{B}}{\mathbf{B}^t\mathbf{B}}. \tag{17}$$

Then from (1)

$$X^1 \leftarrow X - \mathbf{G}\mathbf{S}^t. \tag{18}$$

The computation of the temporary vector $\mathbf{G}$ costs only MN complex multiplications and additions (CMADs) plus a square root (see (3)), while the computation of $\mathbf{S}$ takes only about M CMADs, plus a single division. The outer product $\mathbf{G}\mathbf{S}^t$ costs NM more complex multiplications, which are then subtracted from X (except that the M elements of the first row come for free, which approximately cancels the effort needed to compute $\mathbf{S}$). Thus, the total CMAD count is about $2NM$. By comparison, to multiply X by a general $M \times M$ matrix would have required NM^2 CMADs (plus whatever other CMADs were needed to form the $M \times M$ matrix).

We next construct a Householder matrix that zeros out all but the first two elements of the second row of X^1, while leaving the first row intact.

By a series of orthogonal transformations following this pattern, we ultimately find that the array X^N has the form of an $N \times M$ lower triangular matrix. In such a matrix, the last $M - N$ columns are, of course, entirely zero. Since any column which is entirely zero makes no contribution to the correlation matrix, it may be dropped, giving $\hat{X}$, an $N \times N$ lower triangular matrix with the same correlation matrix as the original data matrix X. A detailed description of this Householder triangulation algorithm is found in [1, pp. 40, 41, 148]. The solution of the least-squares problem using (12) is now very easy (without ever finding R), using the fact that

$$R = \hat{X}\hat{X}^t$$

and that $\hat{X}$ is easy to invert (or that systems involving $\hat{X}$ are easy to solve).

The ith stage of the Householder algorithm, which involves an $(N + 1 - i) \times (M + 1 - i)$ matrix (see (16)–(18)), requires about $2(N + 1 - i)(M + 1 - i)$ CMADs, and thus the full triangulation process requires about $N^2M - N^3/3$ CMADs. Solving for the weight vector $\mathbf{W}$ then requires an additional N^2 CMADs. This final step, common to all the methods discussed in this paper, normally represents only a small percentage of the total CMAD count and is ignored in subsequent cost comparisons.

2.3. Numerical issues in solving linear equations. When we speak of inverting a matrix, like $\hat{X}$ or R, there are considerations of an algorithmic nature—the order of multiplication, addition, division, and data movement, and the count of elementary operations needed. But there is a second kind of consideration, the necessary numerical accuracy, which is related to the wordlength we should use on the processor that performs the inversion. Of course, on a general purpose computer, it is possible to provide multiple-precision subroutines and therefore there is no fundamental limitation to the accuracy with which it is possible to solve a given problem. But in practice, multiple-precision subroutines are necessarily slow in comparison with the machine instructions for add, subtract, multiply, and divide, usually slow by a factor much in excess of the ratio in wordlengths. Thus, double precision will usually take more than twice as long as single precision. If we were to design a special purpose processor for a specific problem, we would prefer to design the arithmetic unit with the shortest wordlength consistent with the fundamental nature of the problem.

What is the minimum wordlength requirement for a specific algorithm? It is well documented that the minimum wordlength required to invert a square matrix is closely linked to its eigenvalue spread (see, for instance, [1] and the references therein). We will not attempt to establish this link rigorously here but will rather motivate it heuristically by means of a simple example. Suppose we wish to numerically invert the correlation matrix

$$R = \begin{bmatrix} 1 & & & \\ & \varepsilon & & \\ & & \ddots & \\ & & & \varepsilon \end{bmatrix}. \tag{19}$$

The eigenvalues of R are ε and 1. If the available precision is insufficient to distinguish between ε and 0 then the stored matrix will be singular and hence noninvertible. We thus have the bound

$$\text{wordlength}_{\text{invert } R} \geqq \log_2 \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right) \tag{20}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are, respectively, the largest and smallest eigenvalues of R, and the notation wordlength$_X$ means the wordlength required to accurately perform task X. We cannot hope to succeed in solving our least-squares problem with a wordlength violating (20) if we elect to proceed by explicitly forming R. It would be desirable to have a wordlength bound directly involving the original data matrix X. Such a bound would be a fundamental limit in that no algorithm could solve the given least-squares problem with less precision than this bound dictates. However, X is more likely to be rectangular than square. Therefore, we may not talk about eigenvalues. For rectangular matrices the singular values play a role similar to the eigenvalues of square matrices. The representation

$$X = USV^t \tag{21}$$

is called the singular value decomposition of X, where U and V are orthonormal matrices and S is of the form

$$\left[\begin{array}{ccc|c} s_1 & & & \\ & s_2 & & 0 \\ & & \ddots & \\ & & s_N & \end{array}\right] \quad \text{for } M \geqq N \tag{22}$$

with real positive elements $s_1 \geqq s_2 \geqq \cdots \geqq s_N \geqq 0$. The s_i are called singular values. The singular values of X are simply related to the eigenvalues of R. R can be expressed as

$$R = XX^t = USV^tVS^tU^t = USS^tU^t. \tag{23}$$

But let

$$SS^t = \begin{bmatrix} s_1^2 & & & \\ & s_2^2 & & \\ & & \ddots & \\ & & & s_N^2 \end{bmatrix} \equiv \Lambda. \tag{24}$$

Λ is a diagonal matrix with nonnegative elements. Thus

$$R = U\Lambda U^t$$

is the eigenvalue-eigenvector expansion of R. This demonstrates that the singular values of the rectangular data matrix X are simply the positive square roots of the eigenvalues of the Hermitian positive definite matrix

$$R = XX^t.$$

The concept of "rank" of a rectangular matrix is normally hard to explain, but, given the singular value decomposition of a matrix, its rank is the same as the number of nonzero singular values. Now, suppose the largest singular value is 1 (we can always guarantee this by an appropriate scaling of the matrix). Then, if the smallest singular value is less than 2^{-B}, where B is the wordlength used to store X, then the matrix X can appear rank deficient and hence the least-squares equations will generally be unsolvable ([1, p. 19]).

The bound we seek is, therefore, given by

$$\text{wordlength}_{\text{find } (XX^t)^{-1} \text{ given } X} \geqq \frac{1}{2} \log_2 \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right) \tag{25}$$

which suggests that a possible factor of two savings is available by avoiding the formation of R.

Because we will be working with rectangular matrices of data, there should be no confusion between singular values of X and eigenvalues of R, because X does not have eigenvalues. Parenthetically, if X is square, it has eigenvalues but they are generally different from its singular values. For a square matrix which is Hermitian and positive definite, singular values and eigenvalues coincide. We give the example of

$$X = \begin{bmatrix} -4 & 6 \\ -6 & 4 \end{bmatrix}$$

which is square but not Hermitian. The eigenvalues of X are $\pm j\sqrt{20}$ and they have no significance to our work. The singular values of X are 2 and 10 and their 5:1 ratio is an indication of the precision needed to invert X. The correlation matrix of X is

$$R = \begin{bmatrix} 52 & 48 \\ 48 & 52 \end{bmatrix}$$

and its eigenvalues are the same as its singular values, namely 4 and 100. The ratio 25:1 is an indication of the precision needed to invert R.

We now consider some numerical examples from an adaptive antenna array problem.

2.4. Adaptive array processing via Householder transformations. The antenna array structure to which we have applied the Householder method is that of a sidelobe canceller (SLC) [2]. We would anticipate similar numerical behavior for other types of arrays (such as fully adaptive arrays). An SLC adaptively suppresses interference by forming a weighted linear combination of the signal from the array's main beam and signals from a set of auxiliary beams as shown in Fig. 1. The weights are chosen to minimize the output power, subject to the constraint that the main beam weight be unity so that the desired signal (which has negligible power in the auxiliaries) remains intact, while as much of the interference as possible is eliminated. For this least-squares problem, the constraint in (11) becomes simply the scalar constraint $w_1 = 1$, and the vector $\mathbf{V}$ in (12) becomes $(1, 0, 0, 0, \cdots, 0)^t$.

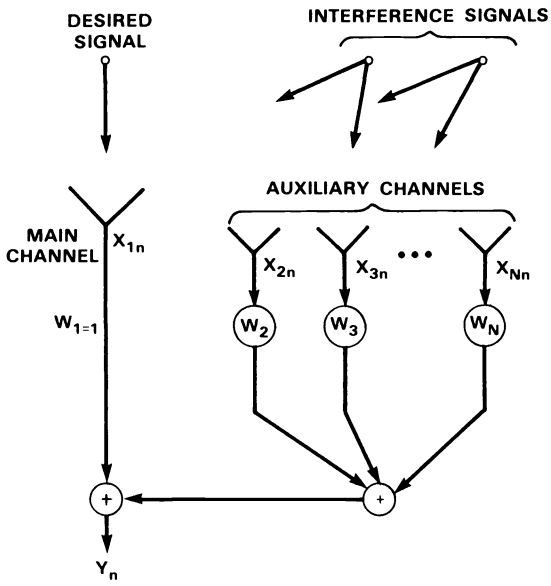


FIG. 1. *A sidelobe canceller adaptive array.*

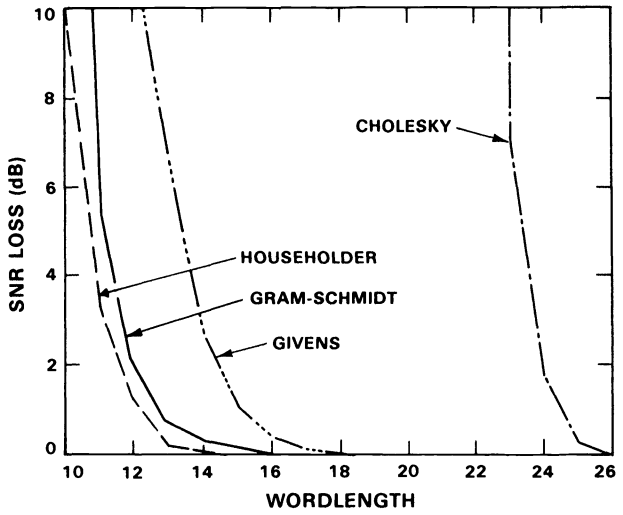


FIG. 2. *Finite wordlength effects of various least-squares algorithms.*

We have chosen for illustration a data matrix of dimension $N = 14$ by $M = 70$. This corresponds to observing interference at 70 sampling instants, on 14 antenna ports simultaneously. The eigenvalue spread for the correlation matrix R for the simulated data which we generated was about 73dB (this spread varies with the interference to receiver-noise ratio).

The loss in output SNR[2] as a function of wordlength (in fixed point arithmetic), averaged over 10 runs, is shown in Fig. 2. We see that the direct inversion of R (by means of the Cholesky method [1]) exhibits a definite threshold phenomenon, with the threshold located near 25 bits (including the sign bit), as predicted by the bound in (24). Since the Cholesky method threshold is close to the wordlength bound specified in (20), we could hope to do no better with any alternate "power domain" method, i.e., one employing the matrix R directly. The Householder method, in contrast, exhibits a much lower threshold, around 13 bits, as predicted by the bound in (25). This algorithm thus has near optimal numerical accuracy, in that no other algorithm can "invert" the data matrix X using less numerical precision. Also shown in Fig. 2 are two other voltage domain algorithms with similar numerical performance, the Givens rotation and the modified Gram–Schmidt [1]. The latter of these has been advocated for use in adaptive arrays in applications where massive computational parallelism is required [3]. Of the three voltage domain methods, the Householder method is the least expensive computationally.

We would like to carry out the Householder transformations using fixed-point arithmetic. We first give a simple argument which shows that this should not be practical, then present a way around the problem by means of a simple, fixed, scaling rule.

The essence of the first Householder transformation is to fold all the energy of the first row of X into its first element. Therefore, in the worst case, when all elements x_{1n} of the first row of X are as large (in energy) as possible, $|x_{11}^1|$ will be $\sqrt{M}$ times larger. In any case, $|x_{11}^1|$ is $\sqrt{M}$ times as large as the root mean square average of the elements in the first row of X. Therefore, we certainly cannot avoid the need to provide for some

[2] The loss is referenced to the signal to noise ratio (SNR) obtained when all calculations are performed with infinite precision. This SNR may be computed by a formula (e.g., [2, (6.8), p. 295]).

scaling for $|x_{11}^1|$. A very similar argument can be made for $|x_{22}^2|$, $|x_{33}^3|$, etc. Of course, two rows might be very similar to one another, so we should allow for energy concentration anywhere in the ith column during the ith Householder transformation. However, these are "final" quantities in the sense that once computed they do not change. Therefore, we scale by the fixed factor $\sqrt{M}$ prior to storing them. But the columns of the X matrix to the right of the ith column are not expected to be magnified by the ith or earlier Householder transformation. These can be retained as fixed point numbers with the original scaling. This argument is not rigorous, but there is the support of the experimental results.

We close this section with an operation count comparison for the Householder and Cholesky approaches. As shown in § 2.2, the Householder method requires $MN^2 - N^3/3$ CMADs. The Cholesky method, in contrast, requires $N^2M/2$ CMADs to compute R from X, and $N^3/6$ CMADs to form the triangular factor of R, for a total of $MN^2/2 + N^3/6$ CMADs. The cost ratio is unity when $M = N$ and increases monotonically, gradually approaching 2:1 in favor of the Cholesky method as M gets large. This rise in computations must be balanced against the attractive factor of two savings in wordlength offered by the Householder approach.

3. Hyperbolic Householder matrices and the updating of least-squares problems.

3.1. Definition and properties of hyperbolic Householder transforms. Let ϕ be a diagonal matrix with diagonal entries $+1$ and -1. Let $\mathbf{x}$ be a complex column vector. We shall call the quantity

$$\mathbf{x}^t\phi\mathbf{x} = \sum_i |x_i|^2\phi_{ii} \tag{26}$$

the "hyperbolic norm" of $\mathbf{x}$, because hyperbolic functions are often characterized by the presence of differences of sums of squares.

We shall call any matrix ψ that satisfies

$$\psi\phi\psi^t = \phi \tag{27}$$

a hypernormal matrix. The justification for this nomenclature is that such matrices preserve the hyperbolic norm of a vector. That is, if $\mathbf{y}^t = \mathbf{x}^t\psi$, then $\mathbf{y}^t\phi\mathbf{y} = \mathbf{x}^t\phi\mathbf{x}$. Generally, a matrix hypernormal with respect to one ϕ matrix will not be hypernormal with respect to another ϕ matrix. Thus, when discussing hypernormality, we must specify the ϕ matrix with respect to which hypernormality has been defined.

We shall call the matrix

$$Q = \phi - \frac{2\mathbf{B}\mathbf{B}^t}{\mathbf{B}^t\phi\mathbf{B}} \tag{28}$$

a hyperbolic Householder matrix.[3]

LEMMA 1. *Q is Hermitian and hypernormal, e.g.,*

$$Q\phi Q^t = \phi. \tag{29}$$

Proof.

$$Q\phi Q^t = \left(\phi - \frac{2\mathbf{B}\mathbf{B}^t}{\mathbf{B}^t\phi\mathbf{B}}\right)\phi\left(\phi - \frac{2\mathbf{B}\mathbf{B}^t}{\mathbf{B}^t\phi\mathbf{B}}\right)$$

$$= \phi^3 - 4\frac{\mathbf{B}\mathbf{B}^t}{\mathbf{B}^t\phi\mathbf{B}} + \frac{4B(\mathbf{B}^t\phi\mathbf{B})B^t}{(\mathbf{B}^t\phi\mathbf{B})^2} = \phi^3 = \phi.$$

[3] Matrices of this type have been previously studied by Bunse-Gerstner in connection with computing eigenvalues [13].

Notice that if $\phi = I$, then Q is an ordinary Householder matrix. Hypernormal matrices, like their simple counterparts, display a great deal of eigenvalue/eigenvector structure. We will discuss this structure in Appendix A, and will make use of our results to establish error bounds. We now show how to construct hyperbolic Householder matrices that zero all but the jth element of a given vector $\mathbf{U}$. In § 3.2, we see how this construction immediately allows us to efficiently insert and delete data from least-squares problems.

We seek to find Q satisfying (28), (29) so that

$$Q\mathbf{U} = \sigma \mathbf{E}_j. \tag{30}$$

The construction follows closely that of the ordinary Householder transform as described in § 2.1.

LEMMA 2. *σ must satisfy the limitation*

$$\mathbf{U}^t \phi \mathbf{U} = |\sigma|^2 \phi_{jj}. \tag{31}$$

Proof. Premultiply each side of (30) on the left by ϕ, and then on the left again by the transpose

$$\mathbf{U}^t Q^t \phi Q \mathbf{U} = \sigma^* \sigma \mathbf{E}_j^t \phi \mathbf{E}_j = |\sigma|^2 \phi_{jj}$$

and replace $Q^t \phi Q$ by ϕ, using (29).

In the following, we assume for simplicity that $\phi_{jj} = 1$. This is always the case for the data deletion/insertion problems that we shall consider. Selecting $\mathbf{B} = \phi\mathbf{U} + \sigma\mathbf{E}_j$, we obtain

$$Q = \phi - \frac{2(\phi\mathbf{U} + \sigma\mathbf{E}_j)(\mathbf{U}^t\phi + \sigma^*\mathbf{E}_j^t)}{(\mathbf{U}^t\phi + \sigma^*\mathbf{E}_j^t)\phi(\phi\mathbf{U} + \sigma\mathbf{E}_j)}. \tag{32}$$

The denominator equals $2\mathbf{U}^t\phi\mathbf{U} + \sigma^* u_j + u_j^* \sigma$, which, again, can be made real by a suitable choice of phase for σ, yielding

$$Q = \phi - \frac{(\phi\mathbf{U} + \sigma\mathbf{E}_j)(\mathbf{U}^t\phi + \sigma^*\mathbf{E}_j^t)}{(\mathbf{U}^t\phi\mathbf{U} + \sigma^* u_j)} \tag{33}$$

with $\sigma = (\pm u_j/|u_j|)\sqrt{\mathbf{U}^t\phi\mathbf{U}}$. Again we elect to use $+u_j$ when forming σ. Hence,

$$\mathbf{U}^t Q = -\sigma^* \mathbf{E}_j^t \tag{34}$$

and we have indeed succeeded in compressing all the hyperbolic energy of $\mathbf{U}$ into its jth entry by means of a transform satisfying the invariance in (29). Again, the formation and application of Q requires about $2NM$ CMADs.

Observe that the structure of this algorithm is identical to that of the conventional Householder algorithm discussed in § 2.1. Consequently, any specialized computer architecture that is well suited to the prior algorithm is well suited to the new algorithm.

There is an interesting connection here with the mathematics of relativity theory. The hyperbolic Householder transforms are a generalization to higher-dimensional complex spaces of the four-dimensional real transform known as the Lorentz transform which arises in the description of spacetime events [5], [6]. The connection derives from the fact that the Lorentz transformation preserves the Minkowski "norm" of two vectors in four-dimensional spacetime. Distance in spacetime is defined in terms of differences of squares, as is the hyperbolic energy which hypernormal matrices preserve.

3.2. Inserting/deleting data with hyperbolic Householder transforms. Suppose we have transformed a given (N by M) matrix X into a correlation equivalent $N \times N$ lower triangular matrix $\hat{X}$ (whether by Householder transformations or by other means). Then

suppose we are given additional data, in the form of an $N \times L$ matrix Y and another $N \times P$ matrix Z formed from some of the columns of the original data matrix X. Y is new data that is added to the problem, and Z is old data which we are deleting from the problem. The correlation matrix, S, for the restated problem, can be computed from the three matrices $\hat{X}$, Y, Z

$$S = \hat{X}\hat{X}^t + YY^t - ZZ^t. \tag{35}$$

Note that if we have formed XX^t, the partial sum ZZ^t was available at some point and could have been saved. Thus $S = XX^t + YY^t - ZZ^t$ and we see some opportunity to reuse old computations. But since we prefer to use an algorithm that avoids the correlation matrices, it is not so obvious how to make use of the fact that $\hat{X}$ is a triangular factor of XX^t.

Our goal is to find a triangular factor of S. We cannot use Householder matrices here because they are orthogonal and hence preserve only positive sums, while (35) contains a difference term. Hyperbolic matrices are, however, well suited for this task. Let ϕ be an $(N + L + P) \times (N + L + P)$ diagonal matrix, with

$$\phi_{ii} = \begin{cases} 1, & i \leqq N+L, \\ -1, & N+L < i \leqq N+L+P. \end{cases} \tag{36}$$

Now form the concatenated matrix $C = [\hat{X}|Y|Z]$. Then

$$C\phi C^t = S. \tag{37}$$

Furthermore, we may replace C by $C\psi$, where ψ is any hypernormal matrix, and we will leave S invariant

$$(C\psi)\phi(\psi^t C^t) = C\phi C^t = S. \tag{38}$$

Now, if a ψ could be found such that $C\psi$ was lower triangular, then our goal would be attained. Such a ψ can be built as a product of N hyperbolic Householder matrices, each, in succession, used to postmultiply C, introducing zeros into row after row of the successive products. This results in the following triangulation algorithm.

HYPERBOLIC HOUSEHOLDER TRIANGULATION ALGORITHM. Given $\hat{X}$, Y, and Z, of dimension $N \times N$, $N \times L$, and $N \times P$, with $\hat{X}$ lower triangular, the following algorithm computes the lower triangular factor $\hat{C}$ of $S = \hat{X}\hat{X}^t + YY^t - ZZ^t$:

BEGIN: Set $C = [\hat{X}|L|Z]$, set ϕ as in (36).
For $i = 1$ to N
  set $\mathbf{U} = (0, \cdots, 0, C_{ii}, 0, \cdots, 0, C_{i,N+1}, C_{i,N+2}, \cdots, C_{i,N+L+P})^t$
  set $\mathbf{B} = \phi\mathbf{U} + \sigma\mathbf{E}_i^t$ where $\sigma = u_i/|u_i|\sqrt{\mathbf{U}^t\phi\mathbf{U}}$
  set $Q = \phi - 2(\mathbf{B}\mathbf{B}^t/\mathbf{B}^t\phi\mathbf{B})$
  set $C = CQ$
Next i
END

Then $\hat{C}$ will be the first N columns of C; the remaining $M - N$ columns of C will be 0.

In writing computer code for this algorithm care can and should be taken to avoid extraneous multiplications and additions by zero. Notice also that multiplication by Q can be performed economically using a method analogous to (16)–(18).

The ith stage in this algorithm requires around $2(L + P + 1)(N - i)$ CMADs, and the entire process $(L + P + 1)N^2$ CMADs. In comparison, the direct formation of S via

(35), and subsequent Cholesky factoring requires about $LN^2/2 + N^3/6$ CMADs if ZZ^t, the correlation matrix of the set to be deleted, is already available (as in sliding window updating for array processing) and $(L + P)N^2/2 + N^3/6$ CMADs if ZZ^t must be computed (as in outlier suppression for robust statistics). For the case $(L + P) \gg N$, the new method is the more costly in CMADs (by a factor of two). In this case the new method may still be attractive because of its nice numerical properties (to be discussed in §§ 3.3, 3.4).

For relatively small update/delete sets such as arise in outlier suppression, the hyperbolic Householder approach simultaneously offers reduced wordlength needs and less computation. Specifically the new method is the less costly when

$$(L+P) < N/3 \quad (\text{for } N \gg 1). \tag{39}$$

For small $(L + P)$, a well-known power domain approach employing Woodbury's identity (see [1], [2]) is more efficient than Cholesky's method. We show in Appendix B that hyperbolic Householder transformation is less costly computationally than this approach whenever $L + P > 2$. It appears that the hyperbolic Householder approach is currently the fastest known method for appending/deleting data in least-squares problems for which (39) is satisfied.

There is a computational savings available if we are continually removing and adding data to S in a systematic fashion, so that the current Y matrix becomes a Z matrix at a later time. Such structured updating occurs, for instance, in the sliding rectangular window application described in § 3.4. Prior to forming C, form $\hat{Y}$, the correlation equivalent lower triangular version of Y. Y can now be replaced by $\hat{Y}$ in (35) since they have the same correlation, and Z can be replaced by its triangular version $\hat{Z}$, formed from an earlier $\hat{Y}$. For $L > N$ the ith stage in the algorithm now requires only $(4i + 2)(N + 1 - i)$ CMADs and the total cost is $LN^2 + N^3/3$, double that of the Cholesky approach. When $L < N$, the savings of the pretriangulation approach is even greater—if $L < N/6$, the hyperbolic Householder approach with pretriangulation uses fewer operations than the Cholesky method. Detailed comparisons are found in § 3.4.

There are at least two published algorithms for deleting (or appending and deleting) data from least-squares problems which are related, albeit remotely, to the hyperbolic Householder transformation method described here. These methods can delete only one element at a time, rather than an entire row.

The first of these, found in [4], employs Givens rotations to remove a single column of real data. The trick is to replace a real column vector $\mathbf{x}$ by $\mathbf{y} = j\mathbf{x}$, so that $\mathbf{y}\mathbf{y}' = -\mathbf{x}\mathbf{x}'$ where $'$ denotes the ordinary transpose without conjugation. We can then employ the complex version of the Givens rotations to rotate $\hat{X}|\mathbf{y}$ into a triangular matrix, which effectively rotates $\mathbf{x}$ out. If the data is already complex, the clever trick no longer works, since in this case, the correlation matrix is formed by conjugate transposition and multiplication of $\mathbf{x}$ by j no longer has any effect on the outer product $\mathbf{x}\mathbf{x}^t$. Furthermore, because Givens rotations are employed rather than Householder reflections, the cost of this approach is higher (by about a factor of two) when more than one column needs to be added or removed.

The second algorithm is described in [1]. This algorithm works on a single column basis as well. The entire set of orthonormal matrices employed to construct the initial triangular factor X is required for this method and the computational cost is high. It does, however, extend to the complex case without significant difficulty.

The above techniques proceed by restructuring the deletion problem so that conventional orthonormal matrices can be employed. But orthonormality is not the meaningful invariant for the deletion problem. By introducing matrices that preserve a meaningful invariant, we are able to significantly reduce the computational cost of "voltage domain" data append/delete operations.

3.3. Numerical issues in data deletion. The numerical precision required in solving the least-squares problem in (10), (11) is accurately predicted by the eigenvalue spread of the correlation matrix R. The precision requirement for the data deletion task is not as easily ascertained. We must be concerned, of course, about the condition of the correlation matrix before updating and about the condition of the correlation matrix after updating, but we must also be concerned about the numerical stability of the transformation from one correlation matrix to the other, even if both correlation matrices are well conditioned. The difference term in (35) can lead to ill-conditioning if, for example, two relatively large nearly equal numbers are encountered.[4] This differencing problem is not eliminated by the voltage domain route. However, the lack of dynamic range doubling does lessen the severity of the ill-conditioning. By monitoring the hyperbolic Householder matrices, we can be alerted to potential numerical problems. Matrix norms are useful for bounding numerical errors. The (L_2) norm of a matrix is defined as

$$\|A\| = \max\left(\frac{\|Ax\|}{\|x\|}\right), \qquad \|x\| \neq 0 \tag{40}$$

where $\|\cdot\|$ is the standard (Euclidean) norm of a vector. We wish to compute

$$\hat{C} = CQ. \tag{41}$$

Instead we obtain

$$\hat{C} + \varepsilon_{\text{out}} = (C + \varepsilon_{\text{in}})Q \tag{42}$$

where ε_{out} is the output error matrix and ε_{in} is the effective equivalent input error matrix. The computation in (42) is unstable if $\|\varepsilon_{\text{in}}\|/\|\varepsilon_{\text{out}}\|$ is large since this would mean the effective input matrix changes significantly even for a small output error. Likewise (41) is unstable if $\|\varepsilon_{\text{out}}\|/\|\varepsilon_{\text{in}}\|$ is large since this implies a large change in the output for small perturbations on the input. Thus we obtain the following index of numerical stability:

$$\kappa \equiv \max\left(\frac{\|\varepsilon_{\text{in}}\|}{\|\varepsilon_{\text{out}}\|}, \frac{\|\varepsilon_{\text{out}}\|}{\|\varepsilon_{\text{in}}\|}\right). \tag{43}$$

Note that $\kappa \geqq 1$. The L_2 norm of a matrix equals the eigenvalue of maximum absolute value [1], i.e.,

$$\|A\| = \max |\lambda_i|. \tag{44}$$

This norm also satisfies the inequality

$$\|AB\| \leqq \|A\|\,\|B\| \tag{45}$$

for any square matrix pair A, B [1]. From (41), (42) we have

$$\|\varepsilon_{\text{out}}\| = \|\varepsilon_{\text{in}}Q\|, \|\varepsilon_{\text{in}}\| = \|\varepsilon_{\text{out}}Q^{-1}\|.$$

Hence from (45)

$$\|\varepsilon_{\text{out}}\| \leqq \|Q\|\,\|\varepsilon_{\text{in}}\|, \qquad \|\varepsilon_{\text{in}}\| \leqq \|\varepsilon_{\text{out}}\|\,\|Q^{-1}\|. \tag{46}$$

From (43), (44), (46) we find

$$\tau = \max(\|Q^{-1}\|, \|Q\|) \geqq \kappa. \tag{47}$$

---

[4] The small difference of large numbers syndrome can be generalized to the case of a nearly rank-deficient matrix difference of large matrices.

For conventional Householder matrices, $\lambda_{\max} = \lambda_{\min} = 1$ and so $\tau = \kappa = 1$. Since the stability index κ can never be less than unity (as can be seen from (43)), Householder matrices possess maximum numerical stability. We find, in Appendix A, that for any hyperbolic Householder matrix,

$$\tau = |\lambda_{\max}| = |1/\lambda_{\min}| = \zeta \pm \sqrt{\zeta^2 - 1} \quad \text{where } \zeta = \frac{\mathbf{B}^t\mathbf{B}}{\mathbf{B}^t\phi\mathbf{B}}. \tag{48}$$

The quantity $\mathbf{B}^t\phi\mathbf{B}$ is automatically computed while constructing Q. We can obtain $\mathbf{B}^t\mathbf{B}$ alongside $\mathbf{B}^t\phi\mathbf{B}$ by means of one extra addition. Thus ζ is found using a single division, and τ with an additional square root and two additions. Observe that τ is monotonically related to ζ. We can thus simply monitor ζ, and deem Q as ill-conditioned if a specified threshold is exceeded. The parameter ζ is the ratio of the conventional energy to the hyperbolic "energy" in $\mathbf{B}$. In practice we have found the bound to be a loose one, although it does reveal ill-conditioning when it arises. This is illustrated by the numerical example considered later.

Numerical issues of data deletion for the related problem of implementing hyperbolic Givens transformations are discussed in [14].

3.4. Square root recursive updating of the adaptive array problem. The weight vector in the adaptive array problem described in § 2.4 needs to be updated regularly to accommodate temporal variations in the interference. This involves solving a new least-squares problem at each update. However, there is a computational savings available by making use of prior work on each update. Algorithms that exploit this potential savings are collectively referred to as recursive least-squares (RLS) algorithms [2], [7]–[9].

The adaptive SLC array considered here consists of auxiliary elements whose outputs are weighted and summed together. As such it is a multichannel system, but of zero order. If tapped delay lines are employed in lieu of simple weights, then we would have a multichannel transversal filter. Fast RLS algorithms exist for these structures [7]–[9], but they all explicitly form a correlation matrix update and hence suffer from dynamic range doubling. The authors are currently investigating the applicability of the hyperbolic Householder transformation to multichannel RLS.

Any updating scheme implies a window on the effective correlation matrix. A natural window choice is a sliding rectangular window. This and other window types are discussed in [9]. In this section a voltage domain algorithm (or square root algorithm) for implementing a sliding rectangular window RLS is developed and its application to a SLC adaptive array is demonstrated.

If data is appended but never deleted from the RLS problem (such as when an exponential or growing rectangular window is employed [7], [8]) then traditional Householder techniques can be employed to construct a square root RLS algorithm. Details of this approach can be found in [4, Chap. 27].

The difficulty with constructing a square root implementation of RLS with a sliding rectangular window lies in the need to subtract as well as add data in the square root domain. Once this has been accomplished it is a simple matter to implement *any* desired window in the square root domain.

The RLS problem we wish to study is described as follows. Suppose we have a sequence of column vectors $\mathbf{X}_i$. From these $\mathbf{X}_i$ we seek to construct the following sequence of weight vectors:

$$\mathbf{W}_i = R_i^{-1}\mathbf{V}, \qquad i = 1, 2, \cdots, \tag{49}$$

$$R_i = \text{Cor}\,([\mathbf{X}_i|\mathbf{X}_{i-1}|\cdots|\mathbf{X}_{i-N+1}]) \tag{50}$$

where Cor $(\cdot)$ is the matrix correlation function, i.e., Cor $(\cdot) = (\cdot)(\cdot)^t$. The vector $\mathbf{V}$ depends on the constraints imposed on the array. For an SLC, $\mathbf{V} = \mathbf{E}_1$. This yields, within a scale factor, the weight vector for the adaptive array described in § 2.4. The weight vector sequence in (49) can be readily modified, if necessary, to accommodate least-squares problems with multiple constraints as described by (7), (8).

Equation (50) establishes a sliding rectangular window of length N. Other window types are similarly derived by selecting a different definition for R_i. For example, a growing exponential window with gain α is defined by

$$R_i = \text{Cor}\,([\mathbf{X}_i|\alpha\mathbf{X}_{i-1}|\cdots|\alpha^{i-2}\mathbf{X}_2|\alpha^{i-1}\mathbf{X}_1]), \qquad i = 1, 2, \cdots. \tag{51}$$

We are interested in the general case of updating a rectangular windowed R_i after every Kth sample. Updating less often than on a sample by sample basis reduces computations and is generally permissible in many interference suppression applications [9]. If only every Kth weight vector is needed, the correlation matrix updating is then given by

$$R_i = R_{i-K} + Y_i Y_i^t - Y_{i-N} Y_{i-N}^t, \tag{52}$$

where $Y_i = [\mathbf{X}_i|\mathbf{X}_{i-1}|\cdots|\mathbf{X}_{i-K+1}]$.

Let the triangular factorization of R_i be $L_i L_i^t$. The square root algorithm bypasses the formation of the R_i by recursing on the triangular factors L_i, using

$$\hat{Y}_i = H([\mathbf{X}_i|\mathbf{X}_{i-1}|\cdots|\mathbf{X}_{i-K+1}]), \tag{53}$$

$$L_i = H_\phi(L_{i-K}|\hat{Y}_i|\hat{Y}_{i-N}) \tag{54}$$

where $H(\cdot)$ denotes ordinary Householder triangulation, and $H_\phi(\cdot)$ denotes the hyperbolic Householder triangulation process as described by the algorithm in § 3.2. The above recursion remains effective if pretriangulation, as effected by the $H(\cdot)$ operation in (53), is omitted. The purpose of this operation is that it saves computations.

Let us consider the computational cost of implementing (53), (54). Two distinct cases arise, depending on which of K and N is largest. When $K > N$, the cost of pretriangulation (53) in CMADs is $KN^2 - N^3/3$. The cost of effecting the ith stage of (54) is $2(2i + 1)(N - i)$, giving $2(N^3/3)$ for all stages. The total cost is then $N^2 + N^3/3$. This is double the cost of a direct Cholesky approach. The Cholesky method benefits from the fact that the correlation matrix of the data to be deleted is already available to be removed from the previous correlation matrix at a cost of only N^2 subtractions.

When $K < N$, the cost of pretriangulating is $NK^2 - K^3/3$. The data matrix in (54) is taller than it is wide. Consequently, in the resulting lower triangular matrix the last $N - K$ rows will not contain zeroed elements. The cost of (54) is then $2(K^3/3)$ (for the upper K rows of the triangular matrix) plus $2K(N - K)^2$ (for the last lower $N - K$ rows). The net cost is then $K^3/3$ at $NK^2 + 2K(N - K)^2$. This is less than the Cholesky method for (approximately) $K < N/6$.

The pretriangular approach is also amenable to parallel processing.

A numerical comparison is offered in Fig. 3. The same array described earlier is used with an update size of $K = N = 14$. The numerically well-conditioned case corresponds to a stationary interference environment, with an eigenvalue spread of 73dB. The hyperbolic Householder method is seen to require slightly more than half the wordlength required by the direct correlation updating method. The ill-conditioned case corresponds to an abrupt change in interference, resulting from the disappearance of the largest source of interference (other sources remain, however), so the eigenvalue spread of the updated correlation matrix is now reduced, to 56dB. Parenthetically, abrupt nonstationarity does

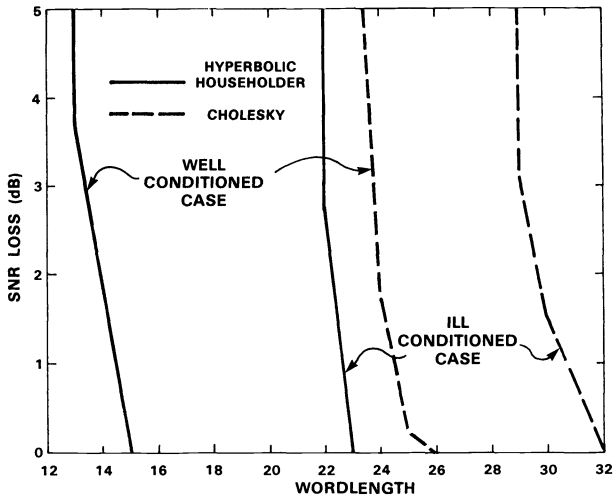


FIG. 3. *Finite wordlength effects for the hyperbolic Householder and Cholesky updating algorithms.*

not always lead to numerical problems—it does so only if it produces the small-difference-of-large-numbers syndrome. Ill-conditioning is found to increase the wordlength needs for both methods, although avoiding the formation of **R** still leads to a significant savings in bits (about 9 bits in the example).

The maximum observed condition number τ for the hyperbolic matrices employed in the stationary case was 7.2, using a 16-bit computation. The condition numbers in the nonstationary case increased noticeably; the largest one became 1119. Condition numbers as high as 10 have been found in cases where no significant performance loss, induced by finite wordlength, was observed. These results suggest that the parameter is a strong indicator of numerical instability. However the "detection threshold" should be chosen to be rather high, exceeding at least 10.

Although we can use hyperbolic Householder transformations to effect updates under a rectangular window, we do not mean to suggest that this process can be carried on indefinitely. By analogy to the simple recursion $y_n = y_{n-1} + x_n - x_{n-K}$, we would expect small computational errors to accumulate from one iteration to the next. For the data of Fig. 3, showing one iteration's result, we have also experimented with multiple iterations. We could see some performance loss after tens of iterations, but with 16 bits this loss was much less than 1dB after 40 iterations, the limit of our experiments.

3.5. Rejecting outliers using hyperbolic Householder matrices. There is a broader class of problems for which we expect hyperbolic Householder transformations to be ideally suited. In least-squares estimation problems we might encounter samples grossly unrepresentative of the data as a whole, such as data incorrectly measured or transmitted. Spurious data, although perhaps difficult to detect directly, often manifests itself as an outlier in the residual. A variety of techniques have been developed for outlier discrimination [10]. Once identified the spurious data can be removed and the least-squares problem resolved with the reduced data set. This iterative least-squares technique has been successfully employed in many applications including the construction of robust all-pole models in spectral estimation [11], [12]. We now see that hyperbolic Householder transformations greatly streamline the re-solution process.

An SLC adaptive array problem is again chosen to illustrate this approach. The interference and the array structure are as described in § 2.4, except that now there is an

occasional additional source of interference (a "blinking" source) that is only rarely present. This blinking source, although present infrequently, can significantly hinder the ability of the antenna to null the steady state interference. This is particularly true if the degrees of freedom in the adaptive array are all required in cancelling the steady state interference so that no further freedom is left to counter the blinking noise source. This is illustrated by Fig. 4. This curve displays the residual output of the adaptive array. The interference-to-receiver-noise ratio is 50dB for both the stationary and the blinking interference. Thus we would expect to achieve nearly a 50dB suppression of interference at the array output under ideal conditions. The sample correlation matrix used to form the nulling weight vector was formed from 100 sample vectors. Three of these sample vectors contain the blinking interference source. The residual is formed from applying the nulling weights to these 100 sample vectors. The three nulled spurious samples are indicated by arrows. As expected they are not well nulled. More significantly, the non-spurious steady state samples are likewise poorly nulled. In fact, the average interference suppression is only 13dB, well below the 47dB suppression obtained with no spurious samples. This is because the spurious samples corrupt the correlation matrix, and consequently the weight vector used to form the residual.

A simple and effective method for avoiding this effect is to reject from the correlation estimate samples that produce large residuals. A rejection threshold must be selected. We chose a threshold of two residual standard deviations. This resulted in all three spurious samples being correctly removed, along with a fourth, good sample. A new weight vector was now computed from the modified correlation matrix. The resulting residual was now much smaller, with a resulting suppression of 47dB. When 100 new samples are nulled with this same weight vector, the 47dB of noise cancellation is still observed.

The above outlier removal technique is well known. More sophisticated schemes are discussed in [10]–[12]. Many of these schemes proceed by monitoring the residual and recomputing a least-squares fit after the undesirable data has been identified and rejected. Hyperbolic Householder matrices are ideally suited for implementing this procedure, regardless of the scheme employed to identify the bad data. For the above example this approach required 16 bits, as opposed to 26 bits for a direct Cholesky factorization

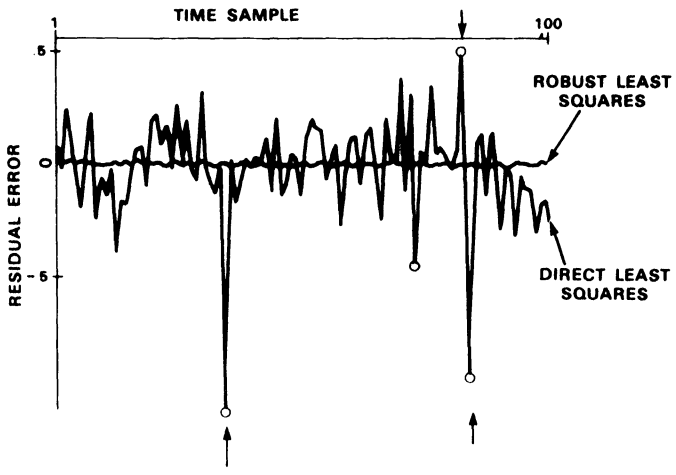


FIG. 4. *The effects of outlier removal on the least-squares residual. The three arrows indicate spurious data. The four circles indicate rejected samples. The robust least-squares residual is formed from resolving the least-squares equations with the rejected samples removed.*

approach. There is a computational savings as well. In fact (see (39)) there is a computational savings whenever the number of rejected samples does not exceed $N/3$.

The above algorithm, including Householder based solution of the initial least-squares problem, is described as follows.

ROBUST LEAST-SQUARES VIA HYPERBOLIC HOUSEHOLDER TRANSFORMATIONS. Given an N parameter linear estimation problem with M sample (data) vectors the following algorithm will compute the robust least-squares solution using a rejection rule of two standard deviations:

$L = H([\mathbf{X}_i, \cdots, \mathbf{X}_M])$
$\mathbf{W} = (LL^t)^{-1}\mathbf{V}$
$\mathbf{e}^t = \mathbf{W}^t[\mathbf{X}_1| \cdots |\mathbf{X}_M]$
$\sigma = 2(\sqrt{e^t e/M})$
$t = 0$
For $i = 1$ to M
  If $|e_i| > \sigma$ then
    $t = t + 1$
    $\mathbf{Y}_t = \mathbf{X}_i$
  $\tilde{L} = H_\phi(L\|\mathbf{Y}_1| \cdots |\mathbf{Y}_t])$
  $\tilde{\mathbf{W}} = (\tilde{L}\tilde{L}^t)^{-1}\mathbf{V}$
Next i

Here $\mathbf{e}$ is the vector of residuals, σ is the threshold, and $\tilde{\mathbf{W}}$ is the resulting robust least-squares estimate. This algorithm is extendible to the case of multiple constraints. Notice that no data is added to L in forming $\tilde{L}$. Thus the Y in (35) is absent. The ϕ matrix, of size $(M + t) \times (M + t)$, has diagonal entries

$$\phi_{ii} = \begin{cases} 1, & i \leqq M, \\ -1, & M < i \leqq M + t. \end{cases}$$

The hyperbolic Householder transforms are applicable to the problem of removing outliers from any linear least-squares problem. However, if the correlation matrix has special structure (such as Hankel or Toeplitz) then direct correlation methods, which can exploit this structure to a computational advantage, may be preferred.

4. Summary. We have defined a class of matrices, the hypernormal matrices, so named because they leave the hyperbolic "norm" of a vector invariant. A matrix from this class (hyperbolic Householder) can always be constructed that zeros all but one element of a specified vector. This construction provides the means for stable and efficient deletion/addition of data from least-squares problems.

The primary motive behind pursuing this method of updating was its lessened sensitivity to rounding errors compared to conventional techniques. When the addition/deletion sets are large, this numerical robustness is obtained at the expense of a small increase in computations. However, when a relatively small fraction of the data set is modified, this method has the added advantage of reduced computations. Since the hyperbolic Householder algorithm has the same structure as its conventional counterpart, parallel matrix processing schemes, such as systolic arrays, can be readily employed to implement it.

Two applications to signal processing problems were considered. In both cases simulations validated the enhanced numerical robustness offered by the new transforms.

Appendix A. The spectral theory of hypernormal matrices. This Appendix addresses the eigenvalue/eigenvector structure of hyperbolic Householder and hypernormal matrices. The following lemmas will be required.

LEMMA 1. *ψ is always nonsingular.*

Proof. Take determinants on both sides of (27):

$$\det(\psi)\det(\phi)\det(\psi^t) = \det(\phi). \tag{A1}$$

Det (ϕ) is nonzero (in fact, it equals ± 1). Dividing both sides of (A1) by this quantity gives det (ψ) det (ψ^t) = 1. Since the determinant is unaltered by matrix transposition, it follows that $|\det(\psi)| = 1$. Hence ψ has a nonzero determinant and is thus nonsingular.

LEMMA 2. *ψ, although generally not symmetric, always satisfies "hyperbolic symmetry," i.e.,*

$$\psi^t\phi\psi = \psi\phi\psi^t. \tag{A2}$$

Proofs.

1. $\psi = \psi$;
2. $\phi^2\psi = \psi$; $\phi^2 = I$;
3. $(\psi\phi\psi^t)\phi\psi = \psi$; $\phi = \psi\phi\psi^t$ by (27);
4. $\psi\phi(\psi^t\phi\psi) = \psi$; associative law;
5. $\psi^t\phi\psi = \phi$; premultiply 4. by $\phi\psi^{-1}$;
6. $\psi^t\phi\psi = \psi\phi\psi^t$; by (27).

THEOREM 1. *The eigenvalues of a hypernormal matrix occur in conjugate reciprocal pairs, i.e., if λ is an eigenvalue, then so is $1/\lambda^*$. Furthermore, the order of λ equals the order of $1/\lambda^*$.*

Proof. Let $\mathbf{V}$ be an eigenvector of ψ. Then

$$(\psi - \lambda I)\mathbf{V} = 0. \tag{A3}$$

Multiply by $\psi^t\phi$. Then from (A3), (27) we have that

$$(\phi - \lambda\psi^t\phi)\mathbf{V} = 0$$

or

$$\psi^t(\phi\mathbf{V}) = (\phi\mathbf{V})/\lambda$$

which upon transposition yields

$$(\phi\mathbf{V})^t\psi = (\phi\mathbf{V})^t/\lambda^*. \tag{A4}$$

Hence, $1/\lambda^*$ is an eigenvalue of ψ. Notice that $(\phi\mathbf{V})^t$ is a row eigenvector of ψ.

Now, suppose that λ has multiplicity k. Two cases arise depending on whether or not its corresponding eigenvectors are linearly independent.

For linearly independent eigenvectors, we can easily establish that λ and $1/\lambda^*$ have equal multiplicity. Indeed, let $\mathbf{V}_1, \cdots, \mathbf{V}_k$ be linearly independent eigenvectors with the eigenvalue λ. Then from (A4), $(\phi\mathbf{V}_1)^t, \cdots, (\phi\mathbf{V}_k)^t$ are row eigenvectors of $1/\lambda^*$, that (since ϕ is square and invertible) are all linearly independent. Reversing the roles of λ, $\mathbf{V}_i$, and $1/\lambda^*$, $\phi\mathbf{V}_i$, we find that λ and $1/\lambda^*$ both must have a multiplicity of exactly k.

The degenerate case of dependent eigenvalues easily follows in a similar fashion. If λ has an order k degeneracy, then there are generalized eigenvectors satisfying

$$(Q - \lambda I)^{i-1}\mathbf{V}_i \neq 0,$$

$$(Q - \lambda I)^i\mathbf{V}_i = 0, \qquad i = 1, \cdots, k.$$

We can make the substitution $\mathbf{U}_i = (Q - \lambda I)^{i-1}\mathbf{V}_i$ whereupon the proof for the non-degenerate case can be applied. □

We now consider the eigenvalue structure of the hyperbolic Householder matrices, defined by (28).

LEMMA 3. *Let ϕ be a diagonal $N \times N$ matrix, with k diagonal elements that are -1, and $N - k$ diagonal elements that are unity. Then any matrix Q given by* (28) *has at least $N - (k + 1)$ eigenvalues equal to* 1, *and at least $k - 1$ eigenvalues equal to -1. (This accounts for all but possibly* 2 *eigenvalues.)*

Proof. Let $Z = Q - I = \phi - I - 2\mathbf{B}\mathbf{B}^t/\mathbf{B}^t\phi\mathbf{B}$. $\phi - I$ is a diagonal matrix of rank k. Since $\mathbf{B}\mathbf{B}^t$ is an outer product of a vector, it has unit rank. Since Z is a sum of a rank k matrix and a rank one matrix, it has a rank of at most $k + 1$, producing an $(N - (k + 1))$th order root at $\lambda = 1$ in the characteristic polynomial $\det|Q - \lambda I|$. Hence, 1 is a multiple eigenvalue of Q with the specified multiplicity. By replacing I by $-I$ in the above argument, we likewise find that at least $k - 1$ eigenvalues of Q are at -1. □

Q is Hermitian, and thus has real eigenvalues [5]. From Theorem 1, the remaining two eigenvalues are thus reciprocals. The two eigenvalues we seek must solve the characteristic polynomial $\det((Q - \lambda I)(Q - \lambda^{-1}I)) = 0$. Let $r = -[\lambda + \lambda^{-1}]$. Then we can rewrite this as $\det((Q + rI)Q + I) = 0$. The required r is easily seen to be $r = 2\mathbf{B}^t\mathbf{B}/\mathbf{B}^t\phi\mathbf{B}$. Upon solving for λ, we establish the following theorem.

THEOREM 2. *Q has $(N - (k + 1))$ of its eigenvalues at* 1, *$k - 1$ at -1, and the remaining two at*

$$\lambda = -\zeta \pm \sqrt{\zeta^2 - 1} \quad \textit{with } \zeta = \frac{\mathbf{B}^t\mathbf{B}}{\mathbf{B}^t\phi\mathbf{B}}.$$

Many of the above lemmas and theorems are natural extensions of well-known properties of orthonormal and Householder matrices. There is, however, one crucial property that hypernormal matrices do not share with their orthonormal counterparts. Unlike orthonormal matrices, hypernormal matrices do not always possess a complete set of linearly independent eigenvectors. Indeed, consider the matrix

$$\psi = \begin{bmatrix} \sqrt{2} & -j \\ 1 & -\sqrt{2}j \end{bmatrix}$$

which is hypernormal with respect to

$$\phi = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

The only eigenvector of ψ is, to within a scale factor, $(1, \exp(j(\pi/4))^t$.

Appendix B. Woodbury's identity and least-squares updating. In this Appendix the computational cost of the Woodbury method of updating is compared to the voltage domain techniques considered in this paper. Bear in mind that the Woodbury identity requires explicit evaluation of the inverse of the correlation matrix R. Hence like Cholesky factoring it is subject to dynamic range doubling, with the attendant poor numerical properties.

Woodbury's identity is a useful tool for finding the inverse of a matrix that has been modified in some minor fashion. For our purposes the following version of this identity will be needed:

$$(R \pm \mathbf{V}\mathbf{V}^t)^{-1} = R^{-1} \mp \frac{R^{-1}\mathbf{V}\mathbf{V}^tR^{-1}}{1 \pm \mathbf{V}^tR^{-1}\mathbf{V}}. \tag{B1}$$

The validity of (B1) is easily established by multiplying both sides by the inverse of the left-hand side. This gives

$$I \stackrel{?}{=} I \pm \mathbf{V}\mathbf{V}^t R^{-1} \mp \frac{\mathbf{V}(1 \pm \mathbf{V}^t R^{-1}\mathbf{V})\mathbf{V}^t R^{-1}}{1 \pm \mathbf{V}^t R^{-1}\mathbf{V}} = I.$$

It requires about $3N^2/2$ CMADs to compute $(R \pm \mathbf{V}\mathbf{V}^t)^{-1}$ given R^{-1}. Thus to update R^{-1} by adding L new vectors and deleting P old vectors requires $3(L + P)N^2/2$ CMADs. The hyperbolic Householder approach requires $(L + P + 1)N^2$ CMADs which is less whenever $L + P > 2$ (when $L + P = 2$ the methods have equal cost).

Acknowledgment. We wish to thank Robert Plemmons of North Carolina State University for providing us with reference [13].

REFERENCES

[1] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[2] R. MONZINGO AND T. MILLER, *Introduction to Adaptive Arrays*, John Wiley, New York, 1980.

[3] C. R. WARD, *Applications of a systolic array to adaptive beamforming*, IEEE Proc., Vol. F, December 1983.

[4] C. LAWSON AND R. HANSEN, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[5] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

[6] W. PAULI, *Theory of Relativity*, Macmillan, New York, 1958, Chap. 24.

[7] J. CIOFFI AND T. KAILATH, *Windowed fast transversal filters adaptive algorithms with normalization*, IEEE Trans. Acoust. Speech Signal Process., 33 (1985), pp. 607–626.

[8] B. PORAT, B. FRIEDLANDER, AND M. MORF, *Square-root covariance ladder algorithms*, IEEE Trans. Automat. Control, 27 (1982), pp. 813–829.

[9] J. CIOFFI, *The block-processing* FTF *adaptive algorithm*, IEEE ICASSP Proceedings, 1985, pp. 1241–1244; ASSP Trans., submitted.

[10] D. HOAGLIN, F. MOSTELLER, AND J. TUKEY, *Understanding Robust and Exploratory Data Analysis*, John Wiley, New York, 1983, Chap. 7.

[11] R. D. MARTIN, *Robust-Resistant Spectral Estimation*, Handbook of Statistics, Vol. 3, D. Brillinger and P. R. Krishnaiah, eds., Elsevier Science, North-Holland, Amsterdam, 1983.

[12] R. MARTIN AND D. THOMSON, *Robust-resistant spectral estimation*, IEEE Proc., September 1982, pp. 1055, 1096.

[13] A. BUNSE-GERSTNER, *An analysis of the* HR *algorithm for computing the eigenvalues of a matrix*, Linear Algebra Appl., 35 (1981), pp. 155–173.

[14] S. ALEXANDER, C. PAN, AND R. PLEMMONS, *Analysis of a recursive least squares hyperbolic rotation algorithm for signal processing*, submitted for publication, 1986.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 2, April 1988

013

NUMERICAL SOLUTION OF THE EIGENVALUE PROBLEM FOR SYMMETRIC RATIONALLY GENERATED TOEPLITZ MATRICES*

WILLIAM F. TRENCH†

Abstract. A numerical method is proposed for finding all eigenvalues of symmetric Toeplitz matrices $T_n = (t_{j-i})_{i,j=1}^n$, where the $\{t_j\}$ are the coefficients in a Laurent expansion of a rational function. Matrices of this kind occur, for example, as covariance matrices of ARMA processes. The technique rests on a representation of the characteristic polynomial as $\det(\lambda I_n - T_n) = W_n G_{0n}(\lambda) G_{1n}$ in which $G_{0n}(\lambda) = 0$ for the eigenvalues of T_n associated with symmetric eigenvectors, $G_{1n}(\lambda) = 0$ for those associated with skew-symmetric eigenvectors, both functions are free of extreme variations, and both can be computed with cost independent of n. It is proposed that root finding techniques be used to compute the zeros of G_{0n} and G_{1n}. Numerical experiments indicate that the method may be useful.

Key words. Toeplitz, rationally generated, eigenvalue, eigenvector

AMS(MOS) subject classifications. 65F15, 15A18, 15A57

1. Introduction.

Let

$$A(z) = a_0 + a_1 z + \cdots + a_q z^q$$

and

$$C(z) = \sum_{j=-p}^{p} c_j z^j,$$

where $a_0, \cdots, a_q$ and $c_{-p}, \cdots, c_0, \cdots, c_p$ are real, $c_j = c_{-j} (1 \leqq j \leqq p)$, $a_q c_p \neq 0$, and $A(z)$ has no zeros in $|z| \leqq 1$. Then the rational function

$$T(z) = \frac{C(z)}{A(z)A(1/z)}$$

has a convergent Laurent expansion

$$T(z) = \sum_{j=-\infty}^{\infty} t_j z^j \tag{1}$$

(with $t_j = t_{-j}$) in an open annulus containing $|z| = 1$.

Here we propose a numerical method for determining the eigenvalues of the symmetric Toeplitz matrices

$$T_n = (t_{j-i})_{i,j=1}^n.$$

Matrices of this kind occur, for example, as covariance matrices of wide-sense stationary autoregressive-moving average time series. (In this setting, $C(z) = B(z)B(1/z)$, with $B(z) = b_0 + b_1 z + \cdots + b_p z^p$.) This is a preliminary report in that further numerical experimentation is required to ascertain whether the method works well for large values of

$$m = \max(p, q); \tag{2}$$

* Received by the editors January 25, 1987; accepted for publication October 1, 1987. This paper was presented at the SIAM Conference on Linear Algebra in Signals, Systems, and Control, which was held in Boston, Massachusetts on August 12–14, 1986.

† Department of Mathematics, Trinity University, San Antonio, Texas 78284.

however, computations already performed with $m = 1, 2$, and 3 indicate that the method can be used very successfully to obtain *all* eigenvalues of T_n at a cost per eigenvalue which depends essentially only on m and is independent of n.

The asymptotic distribution of the eigenvalues of sequences of Toeplitz matrices T_n associated with a convergent Laurent series has been studied extensively. (See, e.g., [8], [13], [19], [20]; there are many other references.) Recently there have been several papers on the spectral structure of symmetric Toeplitz matrices (e.g., [2]–[4], [6], [9]–[11], [15], [16]); however, little has been published on methods for computing the eigenvalues of Toeplitz matrices by methods specifically designed to exploit their simple structure (e.g., [2], [5], [7], [12], [14]). To the author's knowledge, nothing of this kind has been published for rationally generated symmetric Toeplitz matrices, except for the papers of Bini and Capovani [2] and Katai and Rahmy [14], both of which deal only with the case where $A(z) = 1$, so that T_n is banded if $n > q$.

Although it is generally agreed that applying root finding techniques to locate the zeros of its characteristic polynomial is not a good way to find the eigenvalues of a high-order matrix, we believe that this is a viable method for the matrices that we are considering. In order to demonstrate this, we need a theorem proved in [18].

Let $\theta_{-q}, \cdots, \theta_q$ be defined by

$$A(z)A(1/z) = \sum_{j=-q}^{q} \theta_j z^j \tag{3}$$

and define

$$c_j = 0 \quad \text{if } |j| > p, \qquad \theta_j = 0 \quad \text{if } |j| > q.$$

Let $\tau_0, \tau_1, \cdots$ be the Chebyshev polynomials, i.e.,

$$\tau_n(\cos t) = \cos nt. \tag{4}$$

Finally, let

$$p_n(\lambda) = \det [\lambda I_n - T_n]$$

be the characteristic polynomial of T_n, and define

$$C_m(\gamma) = \sum_{j=0}^{q} a_j \cos (n + 2r - 2j - 1)\gamma \tag{5}$$

and

$$S_m(\gamma) = \sum_{j=0}^{q} a_j \sin (n + 2r - 2j - 1)\gamma. \tag{6}$$

Our approach is based on the following theorem, which is proved in [18].

THEOREM 1. *Let λ be such that $c_m - \lambda\theta_m \neq 0$ and the polynomial*

$$P(w; \lambda) = c_0 - \lambda\theta_0 + 2 \sum_{j=1}^{m} (c_j - \lambda\theta_j)\tau_j(w) \tag{7}$$

has m distinct zeros $w_1, \cdots, w_m$ such that

$$w_s \neq 1 \quad \text{or} \quad -1, \qquad 1 \leqq s \leqq m, \tag{8}$$

and let

$$\gamma_s = \frac{1}{2} \cos^{-1} w_s, \qquad 0 \leqq \operatorname{Re}(\gamma_s) \leqq \frac{\pi}{2}. \tag{9}$$

Then

$$p_n(\lambda) = K_n(c_m - \lambda\theta_m)^n F_{0n}(\lambda)F_{1n}(\lambda), \tag{10}$$

where K_n is a constant,

$$F_{0n}(\lambda) = \frac{\det\,[C_{rn}(\gamma_s)]_{r,s=1}^{m}}{\det\,[\cos\,(2r-1)\gamma_s]_{r,s=1}^{m}}, \tag{11}$$

and

$$F_{1n}(\lambda) = \frac{\det\,[S_{rn}(\gamma_s)]_{r,s=1}^{m}}{\det\,[\sin\,(2r-1)\gamma_s]_{r,s=1}^{m}}. \tag{12}$$

Moreover, if $F_{ln}(\lambda) = 0$ ($l = 0$ or 1), then T_n has a λ-eigenvector

$$U = [u_1, \cdots, u_n]^t$$

such that

$$u_{n-i+1} = (-1)^l u_i, \qquad 1 \leqq i \leqq n. \tag{13}$$

We will follow Cantoni and Butler [3] and say that U is *symmetric* if (13) holds with $l = 0$, or *skew-symmetric* if (13) holds with $l = 1$. In [3] it is shown that if T_n is an $n \times n$ real symmetric Toeplitz matrix, then R^n has an orthonormal basis consisting of $n - [n/2]$ symmetric and $[n/2]$ skew-symmetric eigenvectors of T_n. (Here $[x]$ is the integer part of x.) For convenience we will say that the *even spectrum* of T_n consists of eigenvalues with associated symmetric eigenvectors, while the *odd spectrum* consists of eigenvalues with associated skew-symmetric eigenvectors.

Finding the zeros of $P(z; \lambda)$ for a given λ is a nontrivial but tractable (particularly for $1 \leqq m \leqq 4$) problem. Therefore, (10), (11), and (12) *in principle* provide a means for computing $p_n(\lambda)$ for a given λ, with computational cost independent of n, which enters into them only as a parameter. Nevertheless, it is clearly impractical to apply root finding techniques directly to $p_n(\lambda)$ if n is large, simply because a polynomial of high degree can assume tremendous values between its zeros. Fortunately, Theorem 1 provides a way to overcome this difficulty. We will use Theorem 1 to obtain simpler functions which can be evaluated for a given λ with computational cost independent of n, have the same zeros as $F_{0n}(\lambda)$ and $F_{1n}(\lambda)$, and do not vary wildly between their zeros. Root finding techniques can be successfully applied to these functions.

If $m = 1$ in (2), our approach reduces the eigenvalue problem to routine computations and solves it completely (in the numerical sense). We will therefore consider this case separately in § 2. However, some general comments are in order first.

Let

$$f(t) = \frac{C(e^{it})}{A(e^{it})A(e^{-it})}, \qquad -\pi \leqq t \leqq \pi. \tag{14}$$

Then f is real-valued, and $f(t) = f(-t)$; moreover, (3), (4), (7), and (14) imply the identity

$$P(\cos t; f(t)) = 0, \qquad -\pi \leqq t \leqq \pi. \tag{15}$$

It is easily seen that the $\{t_j\}$ in (1) are the Fourier coefficients of f; therefore, if

$$a = \min_{0 \leqq t \leqq \pi} f(t), \qquad b = \max_{0 \leqq t \leqq \pi} f(t), \tag{16}$$

then the eigenvalues of T_n are all in $[a, b]$ for every n (see [8, p. 64]).

For convenience, we will say that the values of λ which do not satisfy the conditions of Theorem 1 are *exceptional points* of $P(\ ;\lambda)$. All other values of λ will be called *ordinary points*. There are at most finitely many exceptional points, and each is of one of the following types: (i) The point $\lambda = c_m/\theta_m$, if $\theta_m \neq 0$. (ii) A value of λ for which the resultant $R(\lambda)$ of the polynomials $P(w;\lambda)$ and $P_w(w;\lambda)$ vanishes; since $R(\lambda)$ is a polynomial in λ, there are only finitely many of these. (iii) The numbers $f(0)$ and $f(\pi)$, since, from (15), $P(1, f(0)) = 0$, $P(-1, f(\pi)) = 0$, which violates (8).

2. The case $m = 1$. If $m = 1$, then $T(z)$ can be written as

$$T(z) = \frac{c_0 + c_1(z + 1/z)}{(1 - \rho z)(1 - \rho/z)}, \tag{17}$$

with $-1 < \rho < 1$. If $\rho = 0$, then T_n is a tridiagonal Toeplitz matrix, and the eigenvalue problem can be solved explicitly (e.g., see [17]). Therefore, we assume that $\rho \neq 0$. We also assume that

$$\rho c_0 + (1 + \rho^2)c_1 = \alpha \neq 0, \tag{18}$$

which guarantees that $T(z)$ does not reduce to a constant, since

$$f'(t) = \frac{-2\alpha \sin t}{(1 - 2\rho \cos t + \rho^2)^2}. \tag{19}$$

Subject to these assumptions, it is straightforward to obtain the expansion (1), with

$$t_j = \frac{c_1\rho^{|j-1|} + c_0\rho^{|j|} + c_1\rho^{|j+1|}}{1 - \rho^2}.$$

Kac, Murdock, and Szegö [13] have considered the special case where

$$c_0 = 1 - \rho^2, \qquad c_1 = 0. \tag{20}$$

The general case can be reduced to this by applying long division to (17), but this would not shorten our presentation. Our results are more detailed than theirs, as we will indicate below.

With T as in (17), (7) becomes

$$P(w;\lambda) = c_0 - \lambda(1 + \rho^2) + 2(c_1 + \lambda\rho)w.$$

Our assumption (18) and its consequence (19) imply that $f(0)$ and $f(\pi)$ are the endpoints of the interval $[a, b]$ defined by (16), and that $c_1 + \lambda\rho \neq 0$, $P(1;\lambda) \neq 0$, and $P(-1;\lambda) \neq 0$ if $a < \lambda < b$. Therefore, Theorem 1 implies that

$$p_n(\lambda) = \frac{K_n(c_1 + \lambda\rho)^n C_n(\gamma) S_n(\gamma)}{\cos\gamma \sin\gamma}, \qquad a < \lambda < b, \tag{21}$$

where K_n is a constant,

$$\gamma = \frac{1}{2}\cos^{-1}\left[\frac{\lambda(1 + \rho^2) - c_0}{2(c_1 + \lambda\rho)}\right], \qquad 0 < \gamma < \frac{\pi}{2},$$

$$C_n(\gamma) = \cos(n + 1)\gamma - \rho\cos(n - 1)\gamma, \tag{22}$$

and

$$S_n(\gamma) = \sin(n + 1)\gamma - \rho\sin(n - 1)\gamma. \tag{23}$$

The formula given in [8] and [13] for the characteristic polynomial of the Kac–Murdock–Szegö matrix

$$T_n = (\rho^{|j-i|})_{i,j=1}^{n},$$

obtained by choosing c_0 and c_1 as in (20), is

$$p_n(\lambda) = \frac{(-\lambda\rho)^n H_n(\gamma)}{(1-\rho^2)\sin 2\gamma}, \tag{24}$$

with

$$H_n(\gamma) = \sin(2n+2)\gamma - 2\rho \sin 2n\gamma + \rho^2 \sin(2n-2)\gamma \tag{25}$$

and

$$\gamma = \frac{1}{2}\cos^{-1}\left[\frac{\lambda(1+\rho^2)-(1-\rho^2)}{2\lambda\rho}\right]. \tag{26}$$

It is observed in [8] that if $H_n(\gamma) = 0$ for some γ in $(0, \pi/2)$, then solving (26) for λ produces an eigenvalue of T_n. It is also shown in [8] that the zeros $\gamma_1, \cdots, \gamma_n$ of H_n satisfy the inequalities

$$0 < \gamma_1 < \frac{\pi}{2n+2} < \gamma_2 < \frac{2\pi}{2n+2} \cdots < \gamma_n < \frac{n\pi}{2n+2}$$

if $0 < \rho < 1$. (The case where $-1 < \rho < 0$ was not considered in [8].)

Numerical computations were not considered in [8], but it is clear that, given such precise information on their locations, $\gamma_1, \cdots, \gamma_n$ can easily be obtained by applying the method of regula falsi to H_n. Therefore, this classical example already illustrates the feasibility of finding the eigenvalues of these matrices by the direct application of root finding techniques, not to $\rho_n(\lambda)$ itself, but to the simple function $H_n(\gamma)$.

Further insight into the eigenvalue problem for this case can be gained by the factorization

$$H_n(\gamma) = 2C_n(\gamma)S_n(\gamma)$$

(cf. (22), (23), (25)), which shows that (21) and (24) are equivalent if (20) holds. Theorem 1 implies that if $C_n(\gamma) = 0$ for some γ in $(0, \pi/2)$ then the quantity

$$\lambda = \frac{c_0 + 2c_1 \cos 2\gamma}{1 - 2\rho \cos 2\gamma + \rho^2} \tag{27}$$

is in the even spectrum of T_n. Also, if $S_n(\gamma) = 0$, then λ in (27) is in the odd spectrum of T_n.

By rewriting (22) and (23) as

$$C_n(\gamma) = (1 - \rho\cos 2\gamma)\cos(n+1)\gamma - \rho \sin 2\gamma \cdot \sin(n+1)\gamma$$

and

$$S_n(\gamma) = (1 - \rho\cos 2\gamma)\sin(n+1)\gamma + \rho \sin 2\gamma \cdot \cos(n+1)\gamma,$$

and then noticing the signs of C_n and S_n at the points

$$x_j = \frac{j\pi}{n+1}, \qquad 0 \leqq j \leqq n - [n/2],$$

and

$$y_j=\frac{(j+\frac{1}{2})\pi}{n+1}, \qquad 0\leqq j\leqq [n/2],$$

it is straightforward to verify that (i) if $0<\rho<1$, then C_n has a zero in (x_{j-1}, y_{j-1}) for each $j=1, \cdots, n-[n/2]$ and S_n has a zero in (y_{j-1}, x_j) for each $j=1, \cdots, [n/2]$; (ii) if $-1<\rho<0$, then C_n has a zero in (y_{j-1}, x_j) for each $j=1, \cdots, n-[n/2]$, and S_n has a zero in (x_j, y_j) for each $j=1, \cdots, [n/2]$.

In either case these zeros are easy to locate by the method of regula falsi. We have written very short BASIC programs to find the zeros and compute the eigenvalues of the Kac–Murdock–Szegö matrices. To illustrate the ease with which they solve the problem, we cite two examples connected with the matrix

$$T_{1000}=(2^{-|j-i|})_{i,j=1}^{n},$$

with computations performed on an IBM PC AT.

(a) With single precision arithmetic (seven significant decimal figures) and requiring the regula falsi iterations to continue until the successive estimates of the zeros of $C_n(\gamma)$ (or $S_n(\gamma)$) agreed in the first six figures, it took 185 seconds to compute the 1000 eigenvalues of T_{1000}.

(b) With double precision arithmetic (16 significant decimal digits) and requiring the regula falsi iterations to continue until the successive estimates of the zeros agreed to 15 places, it took eight minutes to compute the same eigenvalues (of course, to considerably better accuracy than that obtained in (a)).

Since the right side of (27) is a monotonic function of γ in $(0, \pi/2)$ (its derivative, except for a constant, is given by (19) with $t=2\gamma$), our results imply that the odd and even spectra of T_n are interlaced. This has been previously observed by Delsarte and Genin [6].

In the proof of Theorem 1 we gave a general formula ([18, eq. (38)]) for the eigenvectors of rationally generated symmetric matrices. For the special case considered in this section, this formula implies that if $C_n(\gamma)=0$ and λ from (27) is the corresponding eigenvalue, then a corresponding (symmetric) eigenvector is given by $U=[u_1, \cdots, u_n]^t$, with

$$u_i=\cos(n-2i+1)\gamma, \qquad 1\leqq i\leqq n.$$

If $S_n(\gamma)=0$, then

$$u_i=\sin(n-2i+1)\gamma, \qquad 1\leqq i\leqq n,$$

which defines a skew-symmetric eigenvector.

3. The general case ($m\geqq 2$). Now suppose that $m\geqq 2$ and λ is an ordinary point. Then (10), (11), and (12) enable us *in principle* to evaluate $p_n(\lambda)$ with computational cost independent of n, but they suffer from the defect that even though $p_n(\lambda)$ is clearly real if λ is real, (11) and (12) involve complex numbers unless $w_1, \cdots, w_m$ are all in the interval $(-1, 1)$, which is not so in general. Moreover, the tremendous range of values that $p_n(\lambda)$ can assume make it impractical to apply root finding methods directly to $p_n(\lambda)$, or perhaps even to compute it at all.

Fortunately, these problems can be overcome. We will now show that on any subinterval $[a_1, b_1]$ of $[a, b]$ (cf. (16)) containing only ordinary points of $P(\ ;\lambda)$, we can write

$$p_n(\lambda)=W_n(\lambda)G_{0n}(\lambda)G_{1n}(\lambda) \tag{28}$$

where W_n "absorbs" the large variations of $p_n(\lambda)$, but has no zeros on $[a_1, b_1]$ and is therefore of no interest, while G_{0n} and G_{1n} are reasonably computable functions, involving only real quantities, to which root finding methods such as the method of regula falsi or one of its variants can be applied. The zeros of G_{0n} and G_{1n} are, respectively, the elements of the even and odd spectra of T_n which lie in $[a_1, b_1]$.

We need the following lemma, which is established by invoking elementary properties of algebraic functions (see, e.g., [1]) and recalling that P in (7) has real coefficients.

LEMMA 1. *The equation* $P(w; \lambda) = 0$ *defines* m *continuous (in fact, analytic) functions* $w_i = w_i(\lambda)$ $(1 \leqq i \leqq m)$ *on any interval* $[a_1, b_1]$ *consisting entirely of ordinary points of* $P(\ ; \lambda)$. *Moreover, for each* $i = 1, \cdots, m$; (i) $w_i(\lambda)$ *is real for some* λ *in* $[a_1, b_1]$ *if and only if it is real for all such* λ; (ii) *if* w_i *is real-valued, then the functions* $w_i - 1$ *and* $w_i + 1$ *have no zeros on* $[a_1, b_1]$.

This lemma enables us to factor p_n as in (28), where

$$G_{0n}(\lambda) = \det [\tilde{C}_{rn}(\gamma_s)]_{r,s=1}^{m}$$

and

$$G_{1n}(\lambda) = \det [\tilde{S}_{rn}(\gamma_s)]_{r,s=1}^{m}.$$

For each s the definition of $\tilde{C}_{rn}(\gamma_s)$ and $\tilde{S}_{rn}(\gamma_s)$ depends upon whether (i) $-1 < w_s < 1$, (ii) $w_s > 1$, (iii) $w_s < -1$, or (iv) w_s is complex. Lemma 1 implies that exactly one of these conditions holds for all λ in $[a_1, b_1]$.

Case (i). $-1 < w_s < 1$. Here $C_{rn}(\gamma_s)$ and $S_{rn}(\gamma_s)$ are simply linear combinations of (real) sines and cosines, so we let $\tilde{C}_{rn}(\gamma_s) = C_{rn}(\gamma_s)$, and $\tilde{S}_{rn}(\gamma_s) = S_{rn}(\gamma_s)$.

In the remaining cases (9) implies that $\gamma_s = \alpha_s + i\beta_s$, where

$$\cos 2\alpha_s \cosh 2\beta_s = u_s, \quad \sin 2\alpha_s \sinh 2\beta_s = -v_s, \quad 0 \leqq \alpha_s \leqq \frac{\pi}{2} \tag{29}$$

with $w_s = u_s + iv_s$.

Case (ii). $w_s > 1$. Then $w_s = u_s$ and $v_s = 0$ in (29); hence,

$$\alpha_s = 0 \quad \text{and} \quad \beta_s = \tfrac{1}{2} \cosh^{-1} w_s;$$

hence, from (5) and (6),

$$C_{rn}(\gamma_s) = \sum_{j=0}^{q} a_j \cosh (n + 2r - 2j - 1)\beta_s \tag{30}$$

and

$$S_{rn}(\gamma_s) = i \sum_{j=0}^{q} a_j \sinh (n + 2r - 2j - 1)\beta_s. \tag{31}$$

The imaginary unit in the last equation cancels with one which occurs in column s of the denominator in (12). To eliminate large variations, we factor $e^{n\beta_s}/2$ out of the sums in (30) and (31), and define

$$\tilde{C}_{rn}(\gamma_s) = \sum_{j=0}^{q} a_j [e^{(2r-2j-1)\beta_s} + e^{-(2n+2r-2j-1)\beta_s}],$$

$$\tilde{S}_{rn}(\gamma_s) = \sum_{j=0}^{q} a_j [e^{(2r-2j-1)\beta_s} - e^{-(2n+2r-2j-1)\beta_s}].$$

The exponential factor is simply included in $W_n(\lambda)$. Note that $\tilde{C}_{rn}(\gamma_s)$ and $\tilde{S}_{rn}(\gamma_s)$ are bounded for all n. This will also be true in the following cases.

Case (iii). $w_s < -1$. Now

$$\gamma_s = \frac{\pi}{2} + i\beta_s,$$

with

$$\beta_s = \tfrac{1}{2}\cosh^{-1}(-w_s);$$

hence (5) and (6) imply that

$$\pm\sum_{j=0}^{q}(-1)^j a_j \cosh(n+2r-2j-1)\beta_s = \begin{cases} C_{rn}(\gamma_s) & \text{if } n \text{ is odd,} \\ S_{rn}(\gamma_s) & \text{if } n \text{ is even} \end{cases}$$

and

$$\pm i\sum_{j=0}^{q}(-1)^j a_j \sinh(n+2r-2j-1)\beta_s = \begin{cases} S_{rn}(\gamma_s) & \text{if } n \text{ is odd,} \\ C_{rn}(\gamma_s) & \text{if } n \text{ is even.} \end{cases}$$

Therefore, we remove the exponential factor and irrelevant constants as before, and define

$$\sum_{j=0}^{q}(-1)^j a_j[e^{(2r-2j-1)\beta_s} + e^{-(2n+2r-2j-1)\beta_s}] = \begin{cases} \tilde{C}_{rn}(\gamma_s) & \text{if } n \text{ is odd,} \\ \tilde{S}_{rn}(\gamma_s) & \text{if } n \text{ is even} \end{cases}$$

and

$$\sum_{j=0}^{q}(-1)^j a_j[e^{(2r-2j-1)\beta_s} - e^{-(2n+2r-2j-1)\beta_s}] = \begin{cases} \tilde{S}_{rn}(\gamma_s) & \text{if } n \text{ is odd,} \\ \tilde{C}_{rn}(\gamma_s) & \text{if } n \text{ is even.} \end{cases}$$

Case (iv). $w_s = u_s + iv_s$, $v_s \neq 0$. For real λ the coefficients in (7) are real; hence, we may assume without loss of generality that

$$w_{s+1} = u_s - iv_s. \tag{32}$$

If

$$\tau = \cosh^2 2\beta_s, \tag{33}$$

then (29) implies that

$$\frac{u_s^2}{\tau} + \frac{v_s^2}{\tau - 1} = 1,$$

or

$$(\tau-1)u_s^2 + \tau v_s^2 = \tau(\tau-1).$$

Solving this quadratic equation and noting that $\tau > 1$ yields

$$\tau = \tfrac{1}{2}(1 + u_s^2 + v_s^2 + \sqrt{(1+u_s^2+v_s^2)^2 - 4u_s^2}).$$

Now (29) and (33) imply that

$$\alpha_s = \tfrac{1}{2}\cos^{-1}(u_s/\sqrt{\tau})$$

and

$$\beta_s = \tfrac{1}{2} \operatorname{sgn}(-v_s) \cosh^{-1} \sqrt{\tau}. \tag{34}$$

It now follows from (5) and (6) that

$$\begin{aligned} C_{rn}(\gamma_s) = &\sum_{j=0}^{q} a_j \cos(n+2r-2j-1)\alpha_s \cosh(n+2r-2j-1)\beta_s \\ &- i \sum_{j=0}^{q} a_j \sin(n+2r-2j-1)\alpha_s \sinh(n+2r-2j-1)\beta_s \end{aligned} \tag{35}$$

and

$$\begin{aligned} S_{rn}(\gamma_s) = &\sum_{j=0}^{q} a_j \sin(n+2r-2j-1)\alpha_s \cosh(n+2r-2j-1)\beta_s \\ &+ i \sum_{j=0}^{q} a_j \cos(n+2r-2j-1)\alpha_s \sinh(n+2r-2j-1)\beta_s \end{aligned} \tag{36}$$

are the elements in the sth columns of the determinants in the numerators of (11) and (12), respectively. But now (32) and (34) imply that $\gamma_{s+1} = \alpha_s - i\beta_s$, so the elements in the $(s+1)$st columns of these matrices are the conjugates of (35) and (36), i.e.,

$$C_{rn}(\gamma_{s+1}) = \overline{C_{rn}(\gamma_s)}, \qquad S_{rn}(\gamma_{s+1}) = \overline{S_{rn}(\gamma_s)}.$$

This and elementary properties of determinants imply that replacing $C_{rn}(\gamma_s)$ and $C_{rn}(\gamma_{s+1})$ in columns s and $s+1$ of $\det [C_{rn}(\gamma_s)]_{r,s=1}^{n}$ by $\operatorname{Re}(C_{rn}(\gamma_s))$ and $\operatorname{Im}(C_{rn}(\gamma_s))$ simply multiplies this determinant by a purely imaginary constant (which is cancelled by the same constant produced by similar manipulations on the determinant in the denominator of (11)). Following this by factoring out the exponential $e^{n|\beta_s|}$ and other constants leads us to define the elements in column s of $G_{0n}(\lambda)$ by

$$\tilde{C}_{rn}(\gamma_s) = \sum_{j=0}^{q} a_j [e^{(2r-2j-1)|\beta_s|} + e^{-(2n+2r-2j-1)|\beta_s|}] \cos(n+2r-2j-1)\alpha_s$$

and the elements in column $s+1$ by

$$\tilde{C}_{rn}(\gamma_{s+1}) = \sum_{j=0}^{q} a_j [e^{(2r-2j-1)|\beta_s|} - e^{-(2n+2r-2j-1)|\beta_s|}] \sin(n+2r-2j-1)\alpha_s.$$

Similar operations on the real and imaginary parts of (36) lead to the definitions

$$\tilde{S}_{rn}(\gamma_s) = \sum_{j=0}^{q} a_j [e^{(2r-2j-1)|\beta_s|} + e^{-(2n+2r-2j-1)|\beta_s|}] \sin(n+2r-2j-1)\alpha_s$$

and

$$\tilde{S}_{rn}(\gamma_{s+1}) = \sum_{j=0}^{q} a_j [e^{(2r-2j-1)|\beta_s|} - e^{-(2n+2r-2j-1)|\beta_s|}] \cos(n+2r-2j-1)\alpha_s$$

as the entries in columns s and $s+1$ of $G_{1n}(\lambda)$.

4. A proposed procedure for finding all eigenvalues of T_n. Here we assume that $m \geqq 2$, since § 2 reduces the case where $m = 1$ to routine computations. We pro-

pose a procedure for computing all the eigenvalues of T_n. As mentioned earlier, it is known that the spectrum of T_n is contained in the interval $[a, b]$. For simplicity we will assume here that the eigenvalues $\lambda_{1n}, \cdots, \lambda_{nn}$ are distinct, that no exceptional point of $P(\ ; \lambda)$ is an eigenvalue, and that

$$a < \lambda_{1n} < \lambda_{2n} < \cdots < \lambda_{nn} < b.$$

We consider two situations: (I) the eigenvalues of T_{n-1} are already known and satisfy the inequalities

$$a < \lambda_{1,n-1} < \lambda_{2,n-1} < \cdots < \lambda_{n-1,n-1} < b;$$

and (II) the eigenvalues of T_{n-1} are not known.

The first requirement of the procedure is to subdivide $[a, b]$ into closed subintervals $I_1, \cdots, I_k$ with disjoint interiors such that none of the I_j's contains more than one eigenvalue from each of the even and odd spectra. This is very easily done in situation (I); we simply let $k = n$ and

$$I_j = \begin{cases} [a, \lambda_{1,n-1}] & \text{if } j = 1, \\ [\lambda_{j-1,n-1}, \lambda_{j,n-1}] & \text{if } 2 \leqq j \leqq n-1, \\ [\lambda_{n-1,n-1}, b] & \text{if } j = n. \end{cases}$$

Since T_{n-1} is a principal submatrix of T_n, standard separation theorems imply that each of the intervals $I_1, \cdots, I_n$ contains exactly one eigenvalue of T_n.

Obtaining the desired subdivision of $[a, b]$ in situation (II) requires some guesswork, but the guessing is of the educated variety, thanks to the celebrated theorem of Szegö which says that for large n the eigenvalues of T_n are distributed in $[a, b]$ like the ordinates

$$f\left(\frac{j\pi}{n+1}\right), \qquad 1 \leqq j \leqq n. \tag{37}$$

(For a more precise statement of this result see [8] or [20].) Motivated by this, we have used the following procedure to subdivide $[a, b]$: compute the ordinates (37), list them in memory, and construct a new list $g_1, \cdots, g_n$ consisting of these numbers arranged in nondecreasing order. (Since f' cannot have more than $2m - 2$ zeros on $(0, \pi)$, this can be accomplished efficiently, even for large n.) Then define $g_0 = a$ and $g_{n+1} = b$. In the numerical experiments that we have performed with $m = 2$ and 3 the intervals $I_j = [g_{j-1}, g_j]$, $1 \leqq j \leqq n + 1$ usually satisfy our requirements even for small values of n, like $n = 5$ or $n = 10$. (We would expect that the probability of success with this procedure would increase with n, due to the asymptotic nature of Szegö's theorem. It should also be noted that we do not require that no interval contain more than one eigenvalue; because of our factorization of the characteristic polynomial, an interval may contain two eigenvalues, provided that one belongs to the even spectrum of T_n and the other to the odd.) In some cases we missed a few of the eigenvalues. We then used the "brute force" approach of simply dividing all the intervals into k parts (usually with k arbitrarily chosen to be 5). Obviously, this strategy can be improved.

Now we describe the computations performed for each interval in the subdivision. Let $I = [c, d]$, where it is assumed that I does not contain more than one element from each of the odd and even spectra of T_n and that c and d are not eigenvalues of T_n. Suppose first that I contains only ordinary points of $P(\ ; \lambda)$, and let G_{0n} and G_{1n} be the functions defined on I in § 4. If

$$G_{ln}(c)G_{ln}(d) < 0 \tag{38}$$

for $l = 0$ ($l = 1$), then I contains exactly one element from the even (odd) spectrum of T_n, which can be computed by applying the method of regula falsi to G_{qn}. If (38) does not hold for either $l = 0$ or $l = 1$, then I contains no eigenvalues of T_n.

Now suppose that I contains one or more exceptional points. In this case the definition of G_{qn} will in general change on I. Now we simply subdivide I into subintervals whose interiors contain no exceptional points, pick slightly smaller closed subintervals of these which contain no exceptional points (hoping that no eigenvalue actually lies in the small part of I that is excluded in this process), and apply the above procedure to these intervals. This strategy has worked well in all cases considered.

5. Typical numerical experiments for $m = 2$ and $m = 3$. The computations performed so far have been done with BASIC/D (double precision) programs on an IBM PC AT. No attempt has as yet been made to use more sophisticated programming techniques or numerical methods (such as improvements on the method of regula falsi to find the zeros of G_{0n} and G_{1n}); the objective of these computations was simply to ascertain whether there was any hope that this method would work. The results are quite encouraging. The following are typical examples.

Example 1. We took

$$A(z) = \left(1 - \frac{z}{10}\right)\left(1 - \frac{z}{5}\right)$$

and

$$C(z) = 1.5 - 3.5(z + z^{-1}) + (z^2 + z^{-2}).$$

Regula falsi iterations were continued until successive iterates agreed in the first 15 significant decimal digits. The running times to obtain all eigenvalues of T_n were 0:54 (minutes and seconds) with $n = 10$, 3:44 with $n = 50$, 6:40 with $n = 100$, and slightly over 98 minutes with $n = 1000$. (The last required time seems to be longer than we would expect, given the first three. The author does not know the reason for this.)

Example 2. We took

$$A(z) = 1 - .4z - .47z^2 + .21z^3$$

and

$$C(z) = 1 + 2(z + z^{-1}) - (z^2 + z^{-2}) + (z^3 + z^{-3}).$$

The regula falsi iterations were continued until successive iterates agreed to 12 significant decimal digits. The running times required to compute all eigenvalues were 3:58, 12:56, and 22:10 for $n = 10$, 50, and 100, respectively.

We have obtained partial checks on our results. For example, all of our test computations yielded eigenvalue distributions consistent with what we would expect from Szegö's distribution theorem, and in all cases the eigenvalues of T_n separated those of T_{n+1}.

6. Conclusions and further research. Numerical results obtained so far indicate that this method is an efficient way to compute the eigenvalues of high-order symmetric rationally generated Toeplitz matrices with $m = 1$, 2, or 3 in (2). Since the scaling of G_{0n} and G_{1n} makes these functions bounded for all n, there seems to be no reason why the procedure cannot be applied to very high order matrices, particularly since it does not require that the elements of T_n be stored, or even computed. Moreover, it is obvious that much greater computing speeds can be obtained by using more sophisticated programming

methods and/or equipment. We believe that the major problems which need to be overcome in order to apply this method efficiently for larger values of m in (2) are as follows.

(i) Efficient methods must be developed to find the zeros of $P(\ ;\lambda)$ for a given λ. For $m = 1, 2, 3$ we have simply used standard formulas for the zeros of a polynomial in terms of its coefficients. This method is also a possibility for $m = 4$, but root-finding methods are obviously required for $m \geqq 5$. Of course, there are standard methods available for this problem; moreover, if λ_k and λ_{k+1} are successive iterates obtained in a regula falsi procedure, then the zeros $w_1(\lambda_k), \cdots, w_m(\lambda_k)$ should be reasonably good first approximations to $w_1(\lambda_{k+1}), \cdots, w_m(\lambda_{k+1})$.

(ii) An improved version of the method of regula falsi should be devised. This would include the more or less standard procedure of combining it with bisection, which accelerates convergence in certain situations (i.e., where one endpoint would otherwise "stay in the game" for an excessive number of iterations). It is not clear how further improvements could be attained; for example, iterative methods (such as Newton's) which require the computation of derivatives would not be useful, since we have no computable formulas for the derivatives of G_{0n} and G_{1n} in terms of the zeros of $P(\ ;\lambda)$. Moreover, it seems desirable to insist that the iterative method be one that is guaranteed to converge.

(iii) Accurate computation of the $m \times m$ determinants defining G_{0n} and G_{1n} may be difficult for larger values of m. This does not appear to be a problem for $m = 2$ and $m = 3$. In fact, for $m = 3$ we treated this problem in two ways: (a) a simple cofactor expansion and (b) reduction to triangular form with full pivoting. Although the second method would presumably be more accurate for larger m, there was no difference between the results obtained by the two methods for $m = 3$.

Another area of investigation would be to compute the eigenvectors associated with the eigenvalues obtained by this procedure and check the residuals

$$\alpha = \frac{\|T_n X - \lambda X\|}{\|X\|}.$$

This would be a formidable computation if carried out by brute force; however, the formula given in [18] for the eigenvectors of T_n should greatly simplify this calculation.

REFERENCES

[1] L. V. AHLFORS, *Complex Analysis*, 3rd ed., McGraw-Hill, New York, 1979.

[2] D. BINI AND M. CAPOVANI, *Spectral and computational properties of band symmetric Toeplitz matrices*, Linear Algebra Appl., 52 (1983), pp. 99–126.

[3] A. CANTONI AND F. BUTLER, *Eigenvalues and eigenvectors of symmetric centrosymmetric matrices*, Linear Algebra Appl., 13 (1976), pp. 275–288.

[4] G. CYBENKO, *On the eigenstructure of Toeplitz matrices*, IEEE Trans. Acoust. Speech Signal Process., 32 (1984), pp. 918–921.

[5] G. CYBENKO AND C. VAN LOAN, *Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix*, Tr 82-527, Department of Computer Science, Cornell University, Ithaca, NY, 1984.

[6] P. DELSARTE AND Y. GENIN, *Spectral properties of finite Toeplitz matrices*, in Mathematical Theory of Networks and Systems, Proc. MTNS-83 International Symposium, Beer Sheva, Israel, 1983, pp. 194–213.

[7] D. R. FUHRMANN AND B. LIU, *Approximating the eigenvalues of a symmetric Toeplitz matrix*, Proc. 21st Annual Allerton Conference on Communications, Control, and Computing, 1983, pp. 1046–1055.

[8] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, Los Angeles, 1958.

[9] T. N. E. GREVILLE, *Bounds for the eigenvalues of Hermitian Trench matrices*, Proc. 9th Manitoba Conference on Numer. Math. Comput., Utilitas Mathematica Publ., Winnipeg, 1980, pp. 241–256.

[10] F. A. GRUNBAUM, *Toeplitz matrices commuting with tridiagonal matrices*, Linear Algebra Appl., 40 (1981), pp. 25–36.

[11] ———, *Eigenvectors of a Toeplitz matrix: discrete version of prolate spheroidal wave functions*, SIAM J. Algebraic Discrete Methods, 2 (1981), pp. 136–141.

[12] Y. H. HU AND S. Y. KUNG, *Highly concurrent Toeplitz eigensystem solver for high resolution spectral estimation*, Proc. ICASSP 83, 1983, pp. 1422–1425.

[13] M. KAC, W. L. MURDOCK, AND G. SZEGÖ, *On the eigenvalues of certain Hermitian forms*, J. Rational Mech. and Anal., 2 (1953), pp. 767–800.

[14] I. KATAI AND E. RAHMY, *Computation of the eigensystem of symmetric five diagonal Toeplitz matrices*, Ann. Univ. Sci. Budapest. Sect. Comput., 1 (1978), pp. 9–17.

[15] J. MAKHOUL, *On the eigenvectors of symmetric Toeplitz matrices*, IEEE Trans. Acoust. Speech Signal Process., 29 (1981), pp. 868–872.

[16] D. SLEPIAN, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty—V: The discrete case*, Bell System Tech. J., 57 (1978), pp. 1371–1430.

[17] W. F. TRENCH, *On the eigenvalue problem for Toeplitz band matrices*, Linear Algebra Appl., 64 (1985), pp. 199–214.

[18] ———, *Characteristic polynomials of symmetric rationally generated Toeplitz matrices*, Linear and Multilinear Algebra, 21 (1987), pp. 289–296.

[19] H. WIDOM, *On the eigenvalues of certain Hermitian operators*, Trans. Amer. Math. Soc., 88 (1958), pp. 491–522.

[20] ———, *Toeplitz Matrices*, Studies in Real and Complex Analysis, I. I. Hirschmann, Jr., ed., Prentice–Hall, Englewood Cliffs, NJ, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

NECESSARY AND SUFFICIENT CONDITIONS FOR THE EXISTENCE OF LOCAL MATRIX DECOMPOSITIONS*

PAUL D. GADER†

Abstract. Let $D = (V, E)$ be a directed graph with n vertices. We define the notion of a local matrix with respect to D and we show that every $n \times n$ matrix, over the real or complex numbers, can be factored into a product of local matrices with respect to D if and only if D is strongly connected and contains all loops. We discuss the significance of this result with respect to parallel computation of linear transforms on SIMD processor arrays. We observe that the result can be used to associate with certain irreducible $n \times n$ matrices a generating set of the semigroup of all $n \times n$ matrices under matrix multiplication.

Key words. matrix decompositions, parallel processing, linear transforms, directed graphs, irreducible matrices, neural nets

AMS(MOS) subject classifications. 15A23, 65W05, 05C50

1. Introduction.

1.1. Processor arrays and linear transforms. In recent years, there has been a great deal of interest in the use of parallel processing as a means of increasing the computational capabilities of digital computers. Many types of parallel computer architectures exist and various means of categorizing these architectures have been proposed [6], [8], [10]–[12]. In this paper, we shall be concerned with the type of parallel computer called a *processor array*. A processor array consists of a large number of simple processing elements (PEs), all governed by a single controller. Each PE can perform arithmetic and logic operations and generally has a small amount of memory. In addition, each PE can directly access the memory of a small subset of the other PEs by means of some interconnection network. To execute a step of an algorithm, the controller broadcasts a single instruction to all of the PEs along with a mask bit which determines whether or not a given PE will execute the instruction. Those PEs that are to execute the instruction do so simultaneously, each using the data available in their memories or in the memories of their neighbors. Thus, in one step, a subset of the PEs will execute the same instruction, each on (possibly) different data. Machines of this type are called Single Instruction Multiple Data, or SIMD, machines. A popular example of an SIMD processor array is the *mesh-connected array*. The PEs in a mesh-connected array are arranged in a rectangular grid and have nearest neighbor connections. Examples of such processor arrays are the Massively Parallel Processor (MPP), CLIP-4, ICL Distributed Array Processor (DAP) and the Geometric Arithmetic Parallel Processor (GAPP) [5], [7], [10], [11], [14]. Another type of processor array is that based on an n-dimensional hypercube network [3]. It has even been suggested that the Cayley graph of a finite group be used as a model for the interconnection network of a processor array [1], [4], [13].

An important characteristic of these arrays is that each processor can communicate directly with only a small subset of the other processors, its neighbors. Thus, in order to compute efficiently using these arrays, methods should be found for decomposing computations that require information from many processors (global information) into computations that require information from only neighboring processors (local information).

* Received by the editors October 27, 1986; accepted for publication (in revised form) September 8, 1987.
† University of Wisconsin, Oshkosh, Wisconsin 54901; Honeywell Systems and Research Center, Minneapolis, Minnesota 55440; University of Wisconsin, Madison, Wisconsin 53706.

FIG. 1. *A linear array of processors.*

In this paper, we consider the problem of decomposing linear transforms into forms compatible with processor arrays. Let $P_1, P_2, \cdots, P_n$ denote the PEs in a processor array. We assume that PE P_i contains a real or complex number x_i. We establish conditions on the interconnection structure of the processor array which guarantee that every linear transformation of $x = (x_1, x_2, \cdots, x_n)^t$ can be computed using a finite sequence of linear transforms, each of which is local with respect to the interconnection structure. The precise statement of this result is given in Theorem 6, our main result. We remark that these conditions are theoretical, not practical, in nature. They are analogous to conditions for existence of solutions of differential equations. Thus, although our results guarantee the existence of local decompositions, efficient means for computing them must still be developed. Some initial work in this areas has been done [7], [15].

An interesting application of our results is to neural nets. For an excellent introduction to neural nets as well as an extensive list of references see Lippman [9]. Roughly speaking, a neural net can be considered to be a processor array together with an iteration of some function defined on the values associated with the PEs. The structure of a neural net is based on "present understanding of biological nervous systems" [9]. Neural nets are mainly used for pattern recognition and classification, particularly in speech and image problems.

A Hopfield net is one example of a neural net. Let $P_1, P_2, \cdots, P_n$ denote PEs and let $f:\mathbb{R} \to \mathbb{R}$ be the step function

$$f(x) = \begin{cases} 1 & \text{if } x \geqq 0, \\ -1 & \text{if } x < 0. \end{cases}$$

Let $W = (w_{ij})$ be an $n \times n$ matrix over $\mathbb{R}$ and let $v_j(0)$ denote the initial value associated with PE P_j, $j = 1, 2, \cdots, n$. The iteration step is then

$$v_j(t+1) = f\left(\sum_{j=1}^{n} w_{ij} v_j(t) \right), \qquad t = 0, 1, 2, \cdots$$

which is repeated until some stopping condition is reached. The point we would like to emphasize is that a straightforward construction of a Hopfield net requires every PE be connected to every other PE since the matrix W may, in general, be full. Indeed, this has been cited as a disadvantage of the Hopfield net. The results of this paper show that a Hopfield net can be implemented on a processor array with much less connectivity. For example, a Hopfield net can be implemented in a straightforward way on a linear array of processors with nearest neighbor connections, as shown in Fig. 1. This follows from the fact that a consequence of our result, and an earlier related result [15], is that any square matrix can be factored into a product of tridiagonal matrices and tridiagonal matrices are local with respect to linear arrays.

1.2. Mathematical model. Henceforth, all vectors and matrices will be assumed to have complex entries, although the results will also hold true for those having real entries. Let $D = (V, E)$ be a directed graph with vertex set $V = \{v_1, v_2, \cdots, v_n\}$ with $n > 1$ and let $\mathcal{M}_n$ denote the algebra of all $n \times n$ matrices over the complex numbers $\mathbb{C}$. We think of V as the set of PEs. If an ordered pair $(v_i, v_j) \in E$, then we interpret this as meaning that PE v_j can directly access the memory of PE v_i. Since any processor should have access to its own memory, we assume throughout that $(v_i, v_i) \in E$ for $i = 1, 2, \cdots, n$; that is, we assume that D is a directed graph with all loops attached.

DEFINITION 1. We say that a matrix $A = (a_{ij}) \in \mathcal{M}_n$ is *local* with respect to D, or just local if D is understood, if $a_{ij} = 0$ whenever $(v_j, v_i) \notin E$. A *local decomposition* of A (with respect to D) is a decomposition $A = \prod_{i=1}^{k} A_i$ such that A_i is local with respect to D for $i = 1, 2, \cdots, k$. If $A = \prod_{i=1}^{k} A_i$ is a local decomposition, then we say that A has a local decomposition and that the mapping $x \to Ax$ can be executed locally with respect to D.

Note that any function $f: V \to \mathbb{C}$ can be represented by an n-tuple,

$$x = (x_1, x_2, \cdots, x_n)^t.$$

If A has a local decomposition $A = \prod_{i=1}^{k} A_i$, then the linear transform $x \to Ax$ can be computed in k local steps:

$$\begin{aligned} x &\to A_k x = x_1 \\ x_1 &\to A_{k-1} x_1 = x_2 \\ &\vdots \\ x_{k-1} &\to A_1 x_{k-1} = x_k. \end{aligned}$$

Throughout this paper, in particular in Lemmas 3 and 4, we shall show that certain matrices have local decompositions by constructing sequences of local $n \times n$ matrices, $A_1, A_2, \cdots, A_n$, such that $Ax = A_1 A_2 \cdots A_n x$ for $x \in \mathbb{C}^n$.

Let $u, v \in V$. A *path* from u to v is a finite sequence of distinct vertices, except possibly u and v, $u = w_0, w_1, \cdots, w_{k-1}, w_k = v$ such that $(w_i, w_{i+1}) \in E$ for $i = 0, 1, \cdots, k-1$. If there exists a path from u to v, then we say that u is *reachable* from v. We say that D is *strongly connected* if for every $u, v \in V$, u is reachable from v.

We will show that every matrix $A \in \mathcal{M}_n$ has a local decomposition with respect to D if and only if D is strongly connected. Tchuente [15] proved this result in the case where D is a graph. His arguments used results that hold for graphs but not for directed graphs. Our result guarantees that, as long as any PE can access the information in any other PE, at least indirectly, then any linear transform on $\mathbb{C}^m$ with $m \leqq n$ can be computed using a sequence of local, linear transforms.

2. Main theorem.

LEMMA 2. *If every $A \in \mathcal{M}_n$ has a local decomposition with respect to the directed graph D, then D is strongly connected.*

Proof. We prove the contrapositive. Assume that D is not strongly connected and let $v_i, v_j \in V$ such that v_i is not reachable from v_j. Let

$$C_1 = \{u \in V: v_i \text{ is reachable from } u\}$$

and

$$C_2 = \{v \in V: v \text{ is reachable from } v_j\}.$$

By assumption, $C_1 \cap C_2 = \varnothing$. If $B \in \mathcal{M}_n$ is local with respect to D, $v_k \in C_1$ and $v_m \in C_2$, then $b_{km} = 0$, since otherwise v_i would be reachable from v_j. Thus, in particular, the matrix C having $c_{ij} = 1$ and $c_{km} = 0$ if $(k, m) \neq (i, j)$ cannot have a local decomposition with respect to D. □

Assume that D is strongly connected. We show that every permutation matrix has a local decomposition with respect to D. As was observed by Tchuente, if D is a graph, then any permutation can be factored into a product of adjacent transpositions and the problem is trivial. If D is not a graph, then a given permutation may not be expressible as a product of local permutations. Consider, for example, the case of the directed cycle

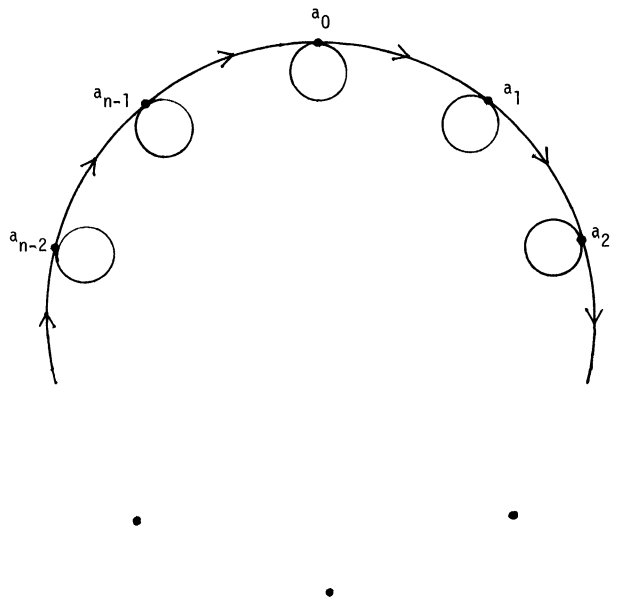


FIG. 2. *The directed cycle* C_n.

with all loops attached C_n shown in Fig. 2. A permutation matrix which is local with respect to C_n is either the identity matrix or has cycle representation $(1\ 2\ \cdots\ n)$. The subgroup of the full symmetric group S_n generated by $(1\ 2\ \cdots\ n)$ is cyclic of order n and is therefore not equal to S_n. Thus, in general, to execute a permutation locally with respect to a directed graph D we must factor the permutation matrix into a product of local matrices which are not necessarily permutation matrices. We now show how this can be done. Note that it is sufficient to show that any transposition of the form

$$(a_0, a_1, \cdots, a_{n-1})^t \to (a_1, a_0, a_2, \cdots, a_{n-1})^t \tag{1}$$

can be executed locally by the following reasoning: If σ is any permutation on $\{0, 1, \cdots, n-1\}$, then we may express σ as a product of transpositions $\sigma = c_1 c_2 \cdots c_n$. If $c = (p, q)$ is a transposition with $q > p$, then we may write

$$\begin{aligned} c = (p,p+1)(p+1,p+2)\cdots(q-2,q-1)(q-1,q) \\ \cdot(q-2,q-1)\cdots(p+1,p+2)(p,p+1). \end{aligned}$$

Thus, if any adjacent transposition, that is, a transposition of the form $(i, i+1)$, can be executed locally with respect to D, then any permutation can be executed locally with respect to D. Since the structure of the graph is independent of the labeling, we may consider, without loss of generality, the transposition (0, 1), thus arriving at (1). Moreover, since D is strongly connected, any two vertices lie on a directed cycle which implies that it is sufficient to consider the case $D = C_n$, the directed cycle with all loops attached. Figure 3 and the example after Lemma 3 are both helpful in the proof of the lemma.

LEMMA 3. *Assume that D is strongly connected with all loops attached. Then every* $n \times n$ *permutation matrix has a local decomposition with respect to D.*

Proof. By the previous remarks, it is sufficient to show that the transposition (1) has a local decomposition with respect to C_n. Let P be the $n \times n$ permutation matrix

a	b	c	d	e	f
a_0	a_0	$\sum_{i=1}^{n-1} a_i$	$\sum_{i=1}^{n-1} a_i$	a_1	a_1
a_1	$a_0 + a_1$	$a_0 + a_1$	$a_0 + a_1$	$a_0 + a_1$	a_0
a_2	$\sum_{i=0}^{2} a_i$	a_2	a_2	a_2	a_2
a_3	$\sum_{i=0}^{3} a_i$	a_3	$a_2 + a_3$	a_3	a_3
.	.	.	.	.	.
.	.	.	.	.	.
a_{n-2}	$\sum_{i=0}^{n-2} a_i$	a_{n-2}	$\sum_{i=2}^{n-2} a_i$	a_{n-2}	a_{n-2}
a_{n-1}	$\sum_{i=0}^{n-1} a_i$	a_{n-1}	$\sum_{i=2}^{n-1} a_i$	a_{n-1}	a_{n-1}

FIG. 3. *Local implementation of permutation* (1).

corresponding to (1), that is,

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \oplus I_{n-2}.$$

Let $a = (a_0, a_1, \cdots, a_{n-1})^t$. We construct a local decomposition of P by constructing a sequence of $n \times n$ local matrices, $Q_1, Q_2, \cdots, Q_{2n+5}$, such that

$$Pa = Q_{2n+5}Q_{2n+4} \cdots Q_2Q_1a.$$

Let $b^{(0)} = a$ and define $b^{(i)} = (b_0^{(i)}, b_1^{(i)}, \cdots, b_{n-1}^{(i)})$, $i = 1, 2, \cdots, n-1$ by

$$b_j^{(i)} = \begin{cases} b_j^{(i-1)} & \text{if } j \neq i, \\ b_{j-1}^{(i-1)} + b_j^{(i-1)} & \text{if } j = i. \end{cases}$$

The matrices Q_i, $i = 1, 2, \cdots, n-1$ with $Q_i b^{(i-1)} = b^{(i)}$ are local. The entries of $b = b^{(n-1)}$ are given in the second column of Fig. 3. Define $c = (c_0, c_1, \cdots, c_{n-1})^t$ by

$$c_j = \begin{cases} b_{n-1} - b_0 & \text{if } j = 0, \\ b_1 & \text{if } j = 1, \\ b_j - b_{j-1} & \text{otherwise.} \end{cases}$$

The matrix Q_n with $Q_n b = c$ is local. The entries of c are given in the third column of Fig. 3. Define $d^{(i)} = (d_0^{(i)}, d_1^{(i)}, \cdots, d_{n-1}^{(i)})^t$, $i = n, n+1, \cdots, 2n+3$ by $d^{(n)} = c$ and for $k = 1, 2, \cdots, n-3$,

$$d_j^{(n+k)} = \begin{cases} d_j^{(n+k-1)} + d_{j-1}^{(n+k-1)} & \text{if } j = k+2, \\ d_j^{(n+k-1)} & \text{otherwise.} \end{cases}$$

The matrices Q_i, $i = n+1, n+2, \cdots, 2n+3$ with $Q_i d^{(i-1)} = d^{(i)}$ are local. The entries of $d = d^{(2n+3)}$ are given in the fourth column of Fig. 3. Define $e = (e_0, e_1, \cdots, e_{n-1})^t$ by

$$e_j = \begin{cases} d_0 - d_{n-1} & \text{if } j = 0, \\ d_1 & \text{if } j = 1, \\ d_2 & \text{if } j = 2, \\ d_j - d_{j-1} & \text{otherwise.} \end{cases}$$

The matrix Q_{2n+4} with $Q_{2n+4} d = e$ is local. The entries of e are given in the fifth column of Fig. 3. Define $f = (f_0, f_1, \cdots, f_{n-1})^t$ by

$$f_j = \begin{cases} e_1 - e_0 & \text{if } j = 1, \\ e_j & \text{otherwise.} \end{cases}$$

The matrix Q_{2n+5} with $Q_{2n+5} e = f$ is local. The entries of f are given in the last column of Fig. 3. By construction, $Pa = (\prod_{i=2n+5}^{1} Q_i)a$ which proves the theorem. □.

Example. Let $n = 5$. Then the sequence of local linear transformations used to execute (1) are given by

$$b^{(0)} = (a_0, a_1, a_2, a_3, a_4),$$

$$b^{(1)} = (a_0, a_0 + a_1, a_2, a_3, a_4),$$

$$b^{(2)} = (a_0, a_0 + a_1, a_0 + a_1 + a_2, a_3, a_4),$$

$$b^{(3)} = \left(a_0, a_0 + a_1, a_0 + a_1 + a_2, \sum_{i=0}^{3} a_i, a_4\right),$$

$$b = b^{(4)} = \left(a_0, a_0 + a_1, a_0 + a_1 + a_2, \sum_{i=0}^{3} a_i, \sum_{i=0}^{4} a_i\right),$$

$$c = \left(\sum_{i=1}^{4} a_i, a_0 + a_1, a_2, a_3, a_4\right),$$

$$d^{(6)} = \left(\sum_{i=1}^{4} a_i, a_0 + a_1, a_2, a_2 + a_3, a_4\right),$$

$$d = d^{(7)} = \left(\sum_{i=1}^{4} a_i, a_0 + a_1, a_2, a_2 + a_3, a_2 + a_3 + a_4\right),$$

$$e = (a_1, a_0 + a_1, a_2, a_3, a_4),$$

$$f = (a_1, a_0, a_2, a_3, a_4).$$

LEMMA 4. *If D is strongly connected, then for $j = 1, 2, \cdots, n-1$ the matrices*

$$M_{j1} = \left[\begin{array}{c|c|c} I_{j-1} & 0 & 0 \\ \hline 0 & \lambda_j & \lambda_{j+1} \cdots \lambda_n \\ \hline 0 & 0 & I_{n-j} \end{array}\right]$$

and

$$M_{j2} = \left[\begin{array}{c|c|c} I_{j-1} & 0 & 0 \\ \hline 0 & \mu_j & 0 \\ \hline & \mu_{j+1} & \\ 0 & \vdots & I_{n-j} \\ & \mu_n & \end{array}\right]$$

have local decompositions with respect to D.

Proof. Fix j between 1 and $n-1$ and note that if $a = (a_1, a_2, \cdots, a_n)^t$ and $b = M_{j1}a = (b_1, b_2, \cdots, b_n)^t$, then $b_i = a_i$ if $i \neq j$ and $b_j = \sum_{k=j}^{n} \lambda_k a_k$. Since D is strongly connected, we can choose m such that $(v_m, v_j) \in E$. Let

$$\Lambda = \operatorname{diag}(1, 1, \cdots, 1, \lambda_j, 1, \cdots, 1)$$

where λ_j is the jth element along the diagonal. For each k such that $j < k \leqq n$, let $P_{k,m}$ denote the permutation matrix corresponding to the transposition (k, m). Let L_k be the matrix having λ_k as the jmth element and zeros elsewhere, and let $S_k = I + L_k$. Since $(v_m, v_j) \in E$, S_k is local. By the previous theorem, each $P_{k,m}$ has a local decomposition. Note that, if $x = (x_1, x_2, \cdots, x_n)^t$, then

$$P_{k,m}S_kP_{k,m}x = (x_1, x_2, \cdots, x_{j-1}, x_j + \lambda_k x_k, x_{j+1}, \cdots, x_n)^t.$$

Hence,

$$M_{j1} = \left(\prod_{k=j+1}^{n} P_{k,m}S_kP_{k,m}\right)\Lambda$$

which shows that M_{j1} has a local decomposition.

Note that if $b = M_{j2}a = (b_1, b_2, \cdots, b_n)^t$, then

$$b_k = \begin{cases} a_k & \text{if } k < j, \\ \mu_j a_j & \text{if } k = j, \\ a_k + \mu_k a_j & \text{if } k > j. \end{cases}$$

Let $\Omega = \operatorname{diag}(1, \cdots, 1, \mu_j, 1, \cdots, 1)$, where μ_j is the jth element along the diagonal. For each k such that $j < k \leqq n$, choose m_k such that $(v_{m_k}, v_j) \in E$, let P_{k,m_k} be the permutation matrix corresponding to the transposition $(k\ m_k)$, let N_k be the matrix having μ_k as the km_kth element and zeros elsewhere, and let $S_k = I + M_k$. Then

$$M_{j2} = \left(\prod_{k=j+1}^{n} P_{k,m_k}S_kP_{k,m_k}\right)\Omega$$

which shows that M_{j2} has a local decomposition. □

Tchuente [15] has shown that if M is any $n \times n$ matrix over $\mathbb{R}$, then there exist $n \times n$ permutation matrices P and Q, constants $\lambda_1, \lambda_2, \cdots, \lambda_n, \mu_1, \mu_2, \cdots, \mu_n$, and an $(n-1) \times (n-1)$ matrix C such that

$$M = P\left[\begin{array}{cc} \lambda_1 & \lambda_2 \cdots \lambda_n \\ 0 & \\ \vdots & I_{n-1} \\ 0 & \end{array}\right]\left[\begin{array}{cc} 1 & 0 \cdots 0 \\ 0 & \\ \vdots & C \\ 0 & \end{array}\right]\left[\begin{array}{cc} \mu_1 & 0 \cdots 0 \\ \mu_2 & \\ \vdots & I_{n-1} \\ \mu_n & \end{array}\right]Q.$$

Although Tchuente states this result over $\mathbb{R}$, an inspection of the proof reveals that it remains valid over $\mathbb{C}$ since only the notions of linear independence of rows and columns of a matrix and solutions of linear systems were used. A straightforward induction argument yields the following lemma.

LEMMA 5. *Let M be any $n \times n$ matrix. There exists $n \times n$ permutation matrices, P_j and Q_j, $n \times n$ matrices, M_{j1} and M_{j2}, of the form given in Lemma 4, for $j = 1, 2, \cdots, n-1$, and a constant, c, such that*

$$M = \left[\prod_{j=1}^{n-1} P_j M_{j1}\right] \operatorname{diag}(1, 1, \cdots, 1, c) \left[\prod_{j=1}^{n-1} M_{j2} Q_j\right].$$

Combining Lemmas 2–5 yields the desired result.

THEOREM 6. *Let $D = (V, E)$ be a direct graph with $n > 1$ vertices. Every $n \times n$ matrix over $\mathbb{R}$ or $\mathbb{C}$ has a local decomposition with respect to D if and only if D is strongly connected with all loops attached.*

3. Irreducible matrices. The notion of a local matrix is closely related to that of an irreducible matrix [2] and therefore Theorem 6 has implications for irreducible matrices which may be useful.

Let A be an $n \times n$ matrix over $\mathbb{C}$. The *digraph of A*, denoted by $\Gamma(A) = (V, E)$, is the directed graph with vertices $\{1, 2, \cdots, n\}$ and arc set $E = \{(i, j): a_{ij} \neq 0\}$. Note that $\Gamma(A)$ is strongly connected if and only if $\Gamma(A^t)$ is. Observe further that if A has all nonzero diagonal elements, then $(i, i) \in E$ for $i = 1, 2, \cdots, n$. We say that A is *irreducible* if $\Gamma(A)$ is strongly connected. Denote by $\mathscr{L}(A)$ the set

$$\mathscr{L}(A) = \{B \in \mathscr{M}_n: a_{ij} = 0 \text{ implies } b_{ij} = 0\}.$$

Note that $B \in \mathscr{L}(A)$ if and only if B^t is local with respect to $\Gamma(A)$. Let $(\mathscr{M}_n, *)$ denote the semigroup of $n \times n$ matrices under matrix multiplication.

THEOREM 7. *The $n \times n$ matrix A is irreducible and has all nonzero diagonal elements if and only if $\mathscr{L}(A)$ generates $(\mathscr{M}_n, *)$.*

Proof. Assume that A is irreducible and has all nonzero diagonal elements and let C be any $n \times n$ matrix. By Theorem 6, C^t has a local decomposition, $C^t = \prod_{i=1}^{k} B_i^t$, with respect to $\Gamma(A)$. Hence, $C = \prod_{i=k}^{1} B_i$ where $B_i \in \mathscr{L}(A)$ for $i = 1, 2, \cdots, k$.

Conversely, if $\mathscr{L}(A)$ generates $(\mathscr{M}_n, *)$, then, given $C \in \mathscr{M}_n$, there are $n \times n$ matrices $B_1, B_2, \cdots, B_k \in \mathscr{L}(A)$ such that $C = \prod_{i=1}^{k} B_i^t$ is a local decomposition of C with respect to $\Gamma(A)$. □

4. Conclusion. We have shown that every $n \times n$ matrix, over $\mathbb{R}$ or $\mathbb{C}$, has a local decomposition with respect to the direct graph D if and only if D is strongly connected with all loops attached. As can be expected for so general a result, the proof, although constructive, is impractical and does not yield methods for obtaining efficient local decompositions.

Acknowledgments. I thank Professor Gerhard X. Ritter for the initial motivation and encouragement to work on these type of problems, Diane Reppert at the University of Wisconsin, Madison, Wisconsin for her invaluable assistance in typing the manuscript, and the referees for helping to improve the clarity of this paper.

REFERENCES

[1] S. B. AKERS AND B. KRISHNAMURTHY, *Group graphs as interconnection networks*, in Digest of Papers, 14th Internat. Conference Fault-Tolerant Computing, June 1984.

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

[3] L. BHUYAN AND D. AGRAWAL, *Generalized hypercube and hyperbus structures for a computer network*, IEEE Trans. Comput., 33 (1983), pp. 323–333.

[4] G. E. CARLSSON, J. E. CRUTHIRDS, H. B. SEXTON AND C. G. WRIGHT, *Interconnection networks based on a generalization of cube-connected cycles*, IEEE Trans. Comput., 8 (1985), pp. 769–779.

[5] M. DUFF, CLIP4—*a large scale integrated circuit array parallel processor*, in 3rd Internat. Joint Conference on Pattern Recognition, Miami, FL, 1976.

[6] M. DUFF AND S. LEVIALDI, EDS., *Languages and Architectures for Image Processing*, Academic Press, New York, 1981.

[7] P. D. GADER, *Tridiagonal decompositions of Fourier matrices with applications to parallel computation of discrete Fourier transforms*, Linear Algebra Appl., 102 (1988), pp. 169–209.

[8] R. HOCKNEY AND C. JESSHOPE, *Parallel Computers: Architecture, Programming, and Algorithms*, Adam Hilger, Bristol, 1981.

[9] R. P. LIPPMAN, *An introduction to computing with neural nets*, IEEE Trans. Acoust. Speech Signal Process., 4 (1987) pp. 4–22.

[10] J. M. ORTEGA AND R. G. VOIGT, *Solution of partial differential equations on vector and parallel computers*, SIAM Rev., 27 (1985), pp. 149–241.

[11] J. L. POTTER, *Image processing on the massively parallel processor*, Computer, 16 (1983), pp. 62–68.

[12] J. T. SCHWARTZ, *A taxonomic table of parallel computers, based on 55 designs*, Tech. Report, Courant Institute of Mathematical Sciences, New York University, New York, 1983.

[13] H. B. SEXTON, *Cayley networks as parallel computer architectures*, in Cooperative Research Associateships tenable at the Naval Ocean Systems Center, National Research Council, Washington, DC, 1986, pp. 12–13.

[14] T. M. SILVERBERG, *The Hough transform on the geometric arithmetic array processor*, in Proc. IEEE Computer Society Workshop on Computer Architecture and Database Management, Miami Beach, FL, November, 1985, pp. 387–394.

[15] M. TCHUENTE, *Parallel calculation of a linear mapping on a computer network*, Linear Algebra Appl., 28 (1979), pp. 223–247.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

002

EXTREMAL PROBLEMS FOR HÖLDER NORMS OF MATRICES AND REALIZATIONS OF LINEAR SYSTEMS*

HARALD K. WIMMER†

Abstract. Let F and G be complex $n \times n$ matrices and $\nu(\cdot)$ be a matrix norm. We consider the functional $\mu(F, G; S) = \nu(FS)\nu(S^{-1}G)$, where S varies over all nonsingular $n \times n$ matrices. For certain singular value norms ν the infimum of μ is determined. An application to realizations of linear systems is given.

Key words. matrix norms, Cauchy–Schwarz inequality, balanced realizations

AMS(MOS) subject classifications. 15A60, 15A23, 93B15

1. Realizations and norms of the Hankel matrix. Let $W \in \mathbb{C}^{p\times q}(z)$ be a given matrix of strictly proper rational functions. If W has the realization

$$W(s) = C(sI - A)^{-1}B, \quad A \in \mathbb{C}^{k\times k}, \quad B \in \mathbb{C}^{k\times q}, \quad C \in \mathbb{C}^{p\times k}, \tag{1.1}$$

then W is the transfer matrix of the linear system

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t), \tag{1.2}$$

and the Markov parameters $W_i \in \mathbb{C}^{p\times q}$ in the expansion $W(s) = \sum_{i>0} W_i z^{-i}$ are given by

$$W_i = CA^{i-1}B, \qquad i = 1, 2, \cdots. \tag{1.3}$$

All the information on W is contained in the block Hankel matrix

$$H = \begin{pmatrix} W_1 & W_2 & \cdots & W_k \\ W_2 & W_3 & \cdots & W_{k+1} \\ \cdot & \cdot & \cdots & \cdot \\ W_k & W_{k+1} & \cdots & W_{2k-1} \end{pmatrix}.$$

If

$$M_o = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{k-1} \end{pmatrix}, \qquad M_c = (B, AB, \cdots, A^{k-1}B)$$

are the observability and controllability matrices of (1.2), then

$$H = M_o M_c.$$

In this paper we are concerned with norms of the Hankel matrix H and its factors M_o and M_c. Let $\sigma_1 \geqq \sigma_2 \geqq \cdots \geqq \sigma_r > 0$ be the singular values of H. Then

$$\|H\|_p = (\sigma_1^p + \cdots + \sigma_1^p)^{1/p} = [\operatorname{Tr}(H^*H)^{p/2}]^{1/p}, \qquad p \geqq 1,$$

and

$$\|H\|_\infty = \sigma_1$$

* Received by the editors December 29, 1986; accepted for publication September 16, 1987.
† Mathematisches Institut, Universität Würzburg, D-8700 Würzburg, Federal Republic of Germany.

are unitarily invariant matrix norms. We focus on $\|H\|_i$, $i = 1, 2, \infty$. The following estimates are crucial for our investigation:

$$\|H\|_i \leqq \|M_o\|_{2i}\|M_c\|_{2i}, \qquad i = 1, 2, \infty.$$

A change of coordinates, $x = S\tilde{x}$, in the state space of (1.2) induces the realization

$$W(s) = \tilde{C}(sI - \tilde{A})^{-1}\tilde{B} \tag{1.4}$$

where $\tilde{A} = S^{-1}AS$, $\tilde{B} = S^{-1}B$, and $\tilde{C} = CS$. We call (1.4) *isomorphic* to the realization (1.1). To (1.4) corresponds the factorization $H = \tilde{M}_o\tilde{M}_c$ with $\tilde{M}_o = M_oS$ and $\tilde{M}_c = S^{-1}M_c$. What happens with the products $\|M_o\|_{2i}\|M_c\|_{2i}$ as S varies over the set of all nonsingular matrices in $\mathbb{C}^{k\times k}$? We will show that $\|H\|_i = \inf \|M_oS\|_{2i}\|S^{-1}M_c\|_{2i}$ in the case $i = 1, 2$ and $\|H\|_\infty = \min \|M_oS\|_\infty\|S^{-1}M_c\|_\infty$. For $i = 1, 2$, we give conditions under which the infima are attained. It will turn out that the stable realizations that actually yield a minimum are those that can be balanced.

Notation. We write $R > 0$ if R is Hermitian and positive definite; $R \geqq 0$ means positive semidefinite. $\mathscr{S}$ is the set of all nonsingular and $\mathscr{P}$ the set of all positive definite matrices in $\mathbb{C}^{k\times k}$. S^{-*} stands for $(S^*)^{-1}$. The largest eigenvalue of a matrix $P \geqq 0$ is denoted by $\lambda_{\max}(P)$. If A is a matrix that has only eigenvalues with negative real parts, then we write $\operatorname{Re}\lambda(A) < 0$.

2. Extremal problems. Let $F \in \mathbb{C}^{p\times k}$ and $G \in \mathbb{C}^{k\times q}$ be given nonzero matrices. We wish to minimize the functionals

$$\mu_i(F, G; S) = \|FS\|_{2i}\|S^{-1}G\|_{2i}, \qquad i = 1, 2, \infty \tag{2.1}$$

where S varies over the set $\mathscr{S}$ of all nonsingular matrices in $\mathbb{C}^{k\times k}$. The following generalization of the Cauchy–Schwarz inequality is surely known. We conjecture that in the case $p = q$ the inequality

$$\|FG\|^2 \leqq \|FF^*\| \quad \|G^*G\|$$

holds for all unitarily invariant norms $\|\cdot\|$ on $\mathbb{C}^{p\times p}$.

LEMMA 2.1. *Suppose $F \in \mathbb{C}^{p\times k}$ and $G \in \mathbb{C}^{k\times q}$, $F \neq 0$. Then*

$$\|FG\|_i \leqq \|F\|_{2i}\|G\|_{2i}, \qquad i = 1, 2, \tag{2.2}$$

with equality if and only if

$$GG^* = \alpha F^*F \tag{2.3}$$

for some real nonnegative α.

Proof. In the case $i = 1$ we have to show that

$$\operatorname{Tr}[(G^*F^*FG)^{1/2}] \leqq (\operatorname{Tr} F^*F)^{1/2}(\operatorname{Tr} G^*G)^{1/2}. \tag{2.4}$$

The traces in (2.4) remain unchanged if F and G are replaced by UF and GV, where U and V are unitary matrices. Hence we can assume without loss of generality

$$FG = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.5}$$

$\Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_r)$, $\sigma_1 \geqq \cdots \geqq \sigma_r > 0$. Let F and G be partitioned corresponding to (2.5):

$$F = \begin{pmatrix} F_1 \\ F_{21} \end{pmatrix}, \qquad G = (G_1, G_{12}).$$

Then $F_1 G_1 = (F_1 G_1)^* = \Sigma > 0$. If we endow the vector space $\mathbb{C}^{k\times r}$ with the inner product $(X, Y) = \operatorname{Tr} X^* Y$, then the Cauchy–Schwarz inequality yields

$$\operatorname{Tr} F_1 G_1 \leqq (\operatorname{Tr} F_1^* F_1)^{1/2} (\operatorname{Tr} G_1 G_1^*)^{1/2}.$$

From $\operatorname{Tr} F^* F = \operatorname{Tr} F_1^* F_1 + \operatorname{Tr} F_{21}^* F_{21} \leqq \operatorname{Tr} F_1^* F_1$ and the analogous inequality for G we obtain (2.4). We have equality if and only if $F_{21} = 0$, $G_{12} = 0$ and $G_1^* = \gamma F_1$. This implies $GG^* = G_1 G_1^* = \|\gamma\|^2 F_1^* F_1 = \|\gamma\|^2 F^* F$. Conversely (2.3) is sufficient for equality.

In the case $i = 2$, the inequality (2.2) is equivalent to

$$\operatorname{Tr} G^* F^* F G \leqq \operatorname{Tr} [(F^* F)^2]^{1/2} \operatorname{Tr} [(G^* G)^2]^{1/2}. \tag{2.6}$$

Because of $\operatorname{Tr} G^* F^* F G = \operatorname{Tr} G G^* F^* F$ we can again use a Cauchy–Schwarz inequality. □

We remark that $\|F\|_{2i} \leqq \|F\|_i$; hence for $i = 1, 2$, the estimates (2.2) are sharper than

$$\|FG\|_i \leqq \|F\|_i \|G\|_i.$$

LEMMA 2.2. *Let $F \in \mathbb{C}^{p\times k}$ and $G \in \mathbb{C}^{k\times q}$. Then*

$$\|FG\|_\infty \leqq \|F\|_\infty \|G\|_\infty \tag{2.7}$$

with equality if and only if the matrices $F^* F$ and GG^* have a common eigenvector which corresponds to their largest eigenvalues.

Proof. Formula (2.7) is known. We check the condition for equality in (2.7). Assume $G \neq 0$. Then

$$\|FG\|_\infty = \max_{y\neq 0} \frac{y^* G^* F^* F G y}{y^* y} = \max_{Gy\neq 0} \frac{y^* G^* F^* F G y}{y^* G^* G y} \frac{y^* G^* G y}{y^* y}$$

$$\leqq \max_{z\neq 0} \frac{z^* F^* F z}{z^* z} \max_{y\neq 0} \frac{y^* G^* G y}{y^* y} = \|F\|_\infty \|G\|_\infty.$$

Put $\alpha = \|F\|_\infty$ and $\beta = \|G\|_\infty$. Then $\alpha^2 = \lambda_{\max} F^* F$ and $\beta^2 = \lambda_{\max} G^* G$. We see that $\|FG\|_\infty = \|F\|_\infty \|G\|_\infty$ if and only if there is a vector $y \in \mathbb{C}^q$ such that $Gy \neq 0$,

$$\beta^2 = \frac{y^* G^* G y}{y^* y} \quad \text{and} \quad \alpha^2 = \frac{y^* G^* F^* F G y}{y^* G^* G y}.$$

Such a y satisfies $G^* G y = \beta^2 y$ and $F^* F G y = \alpha^2 G y$. Then $z = Gy$ is an eigenvector of $F^* F$ which belongs to α^2 and also an eigenvector of GG^* corresponding to β^2. □

The preceding lemmas yield lower bounds for $\mu_i(F, G; S)$ of (2.1), which are in fact infima.

THEOREM 2.3. *Let $F \in \mathbb{C}^{p\times k}$ and $G \in \mathbb{C}^{k\times q}$ be given. Then for $i = 1, 2, \infty$*

$$\|FG\|_i = \inf_{S\in\mathscr{S}} \|FS\|_{2i} \|S^{-1} G\|_{2i}$$

$$= \inf_{R\in\mathscr{P}} \|FRF^*\|_i^{1/2} \|G^* R^{-1} G\|_i^{1/2}. \tag{2.8}$$

Assume $F \neq 0$ and $G \neq 0$. Then there exists an $S \in \mathscr{S}$ such that

$$\|FG\|_\infty = \|FS\|_\infty \|S^{-1} G\|_\infty \tag{2.9}$$

if and only if $FG \neq 0$. For $i = 1$ or $i = 2$ the infimum in (2.8) *is attained, if and only if*

$$\operatorname{rank} F = \operatorname{rank} G = \operatorname{rank} FG. \tag{2.10}$$

For the proof we need some auxiliary results. The construction of an approximating R in (2.8) will hinge on a matrix given by (2.11).

LEMMA 2.4. *Let K and L be two complex $t \times t$ matrices such that K is nonsingular and $KL \geqq 0$. Then there exists a unique Hermitian idempotent E such that*

$$(KL)E = 0 \quad \textit{and} \quad \operatorname{rank}(KL + E) = t.$$

For any $\delta > 0$ the matrix

$$A_\delta = K^{-1}(KL + \delta E)K^{-*} \tag{2.11}$$

is positive definite and

$$L^* A_\delta^{-1} L = KL. \tag{2.12}$$

Proof. Put $M = KL$. If U is unitary such that

$$M = U^* \begin{pmatrix} P & 0 \\ 0 & 0 \end{pmatrix} U, \qquad P > 0,$$

then

$$E = U^* \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} U$$

and the statements of the lemma are readily verified. □

The existence of a minimizing R is related to the following two lemmas.

LEMMA 2.5 [1]. *Let $P \geqq 0$ and $Q \geqq 0$ be of the same size. Then there exists a nonsingular matrix S such that*

$$SPS^* = \begin{pmatrix} I_\tau & 0 \\ 0 & 0 \end{pmatrix}, \qquad S^{-*} Q S^{-1} = \begin{pmatrix} D & 0 \\ 0 & M \end{pmatrix}$$

where $\tau = \operatorname{rank} P$, $D = \operatorname{diag}(d, d_1, \cdots, d_r)$, $\operatorname{rank} D = \operatorname{rank} PQ$, $\operatorname{rank} M = \operatorname{rank} Q - \operatorname{rank} PQ$.

LEMMA 2.6. *For $F \in \mathbb{C}^{p \times k}$ and $G \in \mathbb{C}^{k \times q}$ there exists an $R > 0$ such that*

$$GG^* = RF^*FR \tag{2.13}$$

if and only if

$$\operatorname{rank} F = \operatorname{rank} G = \operatorname{rank} FG. \tag{2.10}$$

Proof. Suppose (2.13) holds for some $R > 0$. Then $\operatorname{rank} F = \operatorname{rank} G$ is obvious. From (2.13) it follows that $FGG^*F^* = (FRF^*)^2$. Therefore

$$\begin{aligned} \operatorname{rank} F &= \operatorname{rank} FRF^* = \operatorname{rank} (FRF^*)^2 \\ &= \operatorname{rank} (FG)(FG)^* = \operatorname{rank} FG. \end{aligned}$$

To prove the converse we first note that $\operatorname{Ker} F^*FGG^* = \operatorname{Ker} FGG^*$. From $\operatorname{rank} FG = \operatorname{rank} G$ it follows that $\operatorname{Ker} FGG^* = \operatorname{Ker} GG^*$. Now put $P = F^*F$ and $Q = GG^*$. Then $\operatorname{rank} P = \operatorname{rank} Q = \operatorname{rank} PQ$. According to Lemma 2.5 there exists an $S \in \mathcal{S}$ such that

$$SPS^* = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \quad S^{-*} Q S^{-1} = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \quad D > 0.$$

Then

$$R = S^* \begin{pmatrix} D^{-1/2} & 0 \\ 0 & I \end{pmatrix} S$$

is a positive definite solution of (2.13). □

Assuming $p = q$ and $FG \geqq 0$, Flanders [3] proved that (2.10) holds if and only if there exists an $R > 0$ such that $G = RF^*$.

Proof of Theorem 2.3. For $i = 1, 2, \infty$ we define

$$\lambda_i(F, G; R) = \|FRF^*\|_i^{1/2} \|G^* R^{-1} G\|_i^{1/2}.$$

We set out to construct a matrix $R_\delta > 0$ such that $\lambda_i(F, G_i; R_\delta) \to \|F\|_i$ with $\delta \to 0$. There is no loss of generality if we assume that $p = q$. Otherwise if $p < q$, we inflate F by zero rows such that

$$\hat{F} = \begin{pmatrix} F \\ 0 \end{pmatrix}$$

is a $q \times k$ matrix. Then $\lambda_i(\hat{F}, G; R) = \lambda_i(F, G; R)$ and $\|\hat{F}G\|_i = \|FG\|_i$. Now choose unitary matrices U, V, W such that

$$\tilde{F} = UFV = \begin{pmatrix} K & 0 \\ 0 & 0 \end{pmatrix}, \quad K \in \mathbb{C}^{t \times t}, \quad \det K \neq 0 \tag{2.14}$$

and $UFGW \geqq 0$. Put $\tilde{G} = V^* GW$ and $\tilde{R} = V^* RV$. Then $\lambda_i(\tilde{F}, \tilde{G}; \tilde{R}) = \lambda_i(F, G; R)$ and $\|FG\|_i = \|\tilde{F}\tilde{G}\|_i$, which allows us to work with $\tilde{F}, \tilde{G}$ instead of the original pair F, G. Thus we can assume

$$F = \begin{pmatrix} K & 0 \\ 0 & 0 \end{pmatrix}$$

as in (2.14) and

$$FG = (FG)^* \geqq 0. \tag{2.15}$$

Let G be partitioned as

$$G = \begin{pmatrix} L & G_{12} \\ G_{21} & G_2 \end{pmatrix}, \qquad L \in \mathbb{C}^{t \times t};$$

then (2.14) and (2.15) imply $KL \geqq 0$ and $G_{12} = 0$. We also note that

$$\|FG\|_i = \|KL\|_i. \tag{2.16}$$

Define $R_\delta \in \mathbb{C}^{k \times k}$ as

$$R_\delta = \begin{pmatrix} A_\delta & 0 \\ 0 & \delta^{-1} I \end{pmatrix}$$

where A_δ is the matrix (2.11). Then

$$FR_\delta F^* = \begin{pmatrix} KL + \delta E & 0 \\ 0 & 0 \end{pmatrix}$$

and (2.12) yields

$$G^*R_\delta^{-1}G=\begin{pmatrix}KL & 0\\ 0 & 0\end{pmatrix}+\delta N$$

where

$$N=G^*G-\begin{pmatrix}L^*L & 0\\ 0 & 0\end{pmatrix}.$$

The following estimates are obvious:

$$\|E\|_i\leqq k,\qquad \|N\|_i\leqq\|N\|_\infty\leqq\|G^*G\|_\infty=\beta^2.$$

Hence

$$\|FR_\delta F^*\|_i\leqq\|KL+\delta E\|_i\leqq\|KL\|_i+\delta k$$

and

$$\|G^*R^{-1}G\|_i\leqq\|KL\|_i+\delta\|N\|_i\leqq\|KL\|_i+\delta\beta^2.$$

From (2.16) and the arithmetic-geometric mean inequality we obtain

$$\lambda_i(F,G;R_\delta)\leqq\|FG\|_i+\tfrac{1}{2}\delta(k+\beta^2),$$

which completes the proof of (2.8).

We now deal with (2.9). Assume $FG\neq 0$. To show that there is an $S\in\mathscr{S}$ such that

$$\|FG\|_\infty=\|F\|_\infty\|G\|_\infty \tag{2.9}$$

we use Lemma 2.5. Let $T\in\mathscr{S}$ be such that

$$TF^*FT^*=\begin{pmatrix}I_\tau & 0\\ 0 & 0\end{pmatrix},\qquad T^{-*}GG^*T^{-1}=\begin{pmatrix}D & 0\\ 0 & KL\end{pmatrix},$$

$D=\operatorname{diag}(d_1,\cdots,d_\tau)$. Then $FG\neq 0$ implies $D\neq 0$ and we can assume $d_1=\max d_\tau$. Choose $\delta>0$ such that $d_1>\lambda_{\max}\,\delta^2KL$ and put

$$S=\begin{pmatrix}I_\tau & 0\\ 0 & \delta^{-1}I\end{pmatrix}T.$$

Then the unit vector $e_1=(1,0,\cdots,0)^T$ is a common eigenvector of SF^*FS^* and $S^{-*}GG^*S^{-1}$ which for both matrices belongs to the largest eigenvalue. According to Lemma 2.2, this implies (2.9).

In the cases $i=1,2$ we see from Lemma 2.1 that

$$\|FG\|_i=\|FS\|_{2i}\|S^{-1}G\|_{2i}$$

if and only if

$$S^{-1}GG^*S^{-*}=\alpha S^*F^*FS,\qquad \alpha>0$$

or

$$GG^*=RF^*FR$$

with $R=\alpha^{1/2}SS^*>0$. From Lemma 2.6 we know that (2.13) is equivalent to the rank condition (2.10). □

A problem in electric circuit theory led Flanders [3], [4] to a functional on $\mathscr{P}$ which involves an arithmetic mean of traces. We phrase his result in our notation.

THEOREM 2.7 [3]. *Let $F, G^* \in \mathbb{C}^{p\times k}$ be given. For $R > 0$, $R \in \mathbb{C}^{k\times k}$ define*

$$f(F,G;R) = \tfrac{1}{2}(\|FRF^*\|_1 + \|G^*R^{-1}G\|_1).$$

Then

$$\text{(2.17)}\quad \begin{aligned} &\text{(a)} \quad \|FG\|_1 = \inf_{R\in\mathscr{P}} f(F,G;R);\\ &\text{(b)} \quad f(F,G;R_0) = \|FG\|_1, \end{aligned}$$

for some $R_0 > 0$ if and only if (2.10) *holds.*

The results for $i = 1$ in (2.9) can be derived from (2.17) and Lemma 2.1. We point out that in our proof of (2.9) the construction of an approximating R_δ is straightforward and works for $i = 1, 2, \infty$.

3. The rank condition and balanced realizations. We return to realizations and the factorization $H = M_oM_c$, $H \neq 0$. With $F = M_o$ and $G = M_c$ Theorem 2.3 settles the problems raised in § 1. The fact that M_o and M_c are the observability and controllability matrices of the linear system (1.2) leads to a different interpretation of the rank condition (2.10). It will be shown that (2.10) holds if and only if each controllable mode of (1.2) is also observable and vice versa. For asymptotically stable systems this means that there exists a state space transformation such that the transformed system has a balanced realization.

THEOREM 3.1. *Let the realization*

$$\text{(1.1)}\qquad W(s) = C(sI-A)^{-1}B,$$

$W \neq 0$, be given.

(a) *Then there exists a realization*

$$\text{(1.4)}\qquad W(s) = \tilde{C}(sI-\tilde{A})^{-1}\tilde{B}$$

isomorphic to (1.1) *with observability matrix $\tilde{M}_o$ and controllability matrix $\tilde{M}_c$ such that*

$$\|H\|_\infty = \|\tilde{M}_o\|_\infty\|\tilde{M}_c\|_\infty.$$

(b) *In the case $i = 1, 2$ there exists for any $\varepsilon > 0$ an isomorphic realization with observability and controllability matrices $\tilde{M}_{o\varepsilon}$ and $\tilde{M}_{c\varepsilon}$ such that*

$$\|\tilde{M}_{o\varepsilon}\|_{2i}\|\tilde{M}_{c\varepsilon}\|_{2i} < \|H\|_i + \varepsilon.$$

There is an isomorphic realization (1.4) *such that*

$$\|H\|_i = \|\tilde{M}_o\|_{2i}\|\tilde{M}_c\|_{2i}, \qquad i = 1,2,$$

if and only if

$$\text{(3.1)}\qquad \operatorname{rank} M_o = \operatorname{rank} M_c = \operatorname{rank} H.$$

(c) *The condition* (3.1) *holds if and only if there is a nonsingular T such that*

$$\text{(3.2)}\qquad T^{-1}AT = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \quad T^{-1}B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}, \quad CT = (C_1, 0)$$

where the pair (A_1, B_1) is controllable and (A_1, C_1) is observable.

Proof. Only (c) has to be proved. The other statements are contained in Theorem 2.3. According to the decomposition theorem for linear systems [5], [7] there is a non-

singular T such that

$$T^{-1}AT=\begin{pmatrix} A_{11} & 0 & A_{13} & 0 \\ A_{21} & A_{22} & A_{23} & A_{24} \\ 0 & 0 & A_{33} & A_{34} \\ 0 & 0 & A_{43} & A_{44} \end{pmatrix}$$

$$T^{-1}B=\begin{pmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{pmatrix}, \qquad CT=(C_1,0,C_3,0), \tag{3.3}$$

where the pair

$$\begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}, \qquad \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$$

is controllable, and

$$\begin{pmatrix} A_{11} & A_{13} \\ 0 & A_{33} \end{pmatrix}, \qquad (C_1, C_3)$$

is observable. Let A_{ij} be of order $k_i \times k_j$. Then rank $M_c = k_{11} + k_{22}$, rank $M_o = k_{11} + k_{33}$ and rank $H = k_{11}$. Hence (3.1) is equivalent to $k_{22} = k_{33} = 0$ and (3.3) reduces to (3.2). □

DEFINITION 3.2 [8]. Let

$$W(s) = C(sI-A)^{-1}B \tag{1.1}$$

be a realization with Re $\lambda(A) < 0$ and let the controllability and observability Gramian of (1.1) be defined as

$$P=\int_0^\infty e^{At}BB^*e^{A^*t}\,dt, \qquad Q=\int_0^\infty e^{A^*t}C^*Ce^{At}\,dt.$$

The realization (1.1) is called *balanced* if $P = Q = \Sigma$ where Σ is a diagonal matrix. We say that (1.1) can be balanced if it is isomorphic to a balanced realization.

The matrices P and M_c, respectively, Q and M_o, are related to each other.

LEMMA 3.3 (see, e.g., [2, p. 79]). *Assume* Re $\lambda(A) < 0$. *Then*

$$P = M_c R M_c^*, \qquad Q = M_o^* R M_o \tag{3.4}$$

for some $R \geqq 0$ *and*

$$\text{rank } P = \text{rank } M_c, \qquad \text{rank } Q = \text{rank } M_o. \tag{3.5}$$

THEOREM 3.4. *A realization* $W(s) = C(sI - A)^{-1}B$ *with* Re $\lambda(A) > 0$ *can be balanced if and only if*

$$\text{rank } M_o = \text{rank } M_c = \text{rank } H \tag{3.1}$$

or equivalently (3.2) *holds.*

Proof. If there is a T such that (3.2), then

$$P = T\begin{pmatrix} P_1 & 0 \\ 0 & 0 \end{pmatrix}T^*, \qquad P_i = \int_0^\infty e^{A_{1i}t}B_1B_1^*e^{A_{1i}^*t}\,dt$$

and

$$Q = T^{-*}\begin{pmatrix} Q_1 & 0 \\ 0 & 0 \end{pmatrix} T^{-1}, \qquad Q_1 = \int_0^\infty e^{A_1^* t} C_1^* C_1 e^{A_1 t}\, dt.$$

As $W_1(s) = C_1(sI - A_1)^{-1}B_1$ is a minimal realization and $\operatorname{Re} \lambda(A_1) > 0$ there is a T_1 (see, e.g., [6, p. 1129]) such that $T_1 P_1 T_1^* = T_1^{-*} Q_1 T_1^{-1} = \Sigma_1$ with a diagonal Σ_1. Hence $x = S\tilde{x}$ with

$$S = T\begin{pmatrix} T_1 & 0 \\ 0 & I \end{pmatrix}$$

is a balancing transformation.

Conversely let us assume that (1.1) can be balanced. Because of (3.4) we have $P^2 = QP = Q^2 = M_o^* RHRM_c^*$ which implies rank P = rank $Q \leqq$ rank H. On the other hand, rank $H \leqq$ rank M_α, $\alpha = o, c$, and (3.5) yields (3.1). □

REFERENCES

[1] T. W. ANDERSON AND I. OLKIN, *An extremal problem for positive definite matrices*, Linear and Multilinear Algebra, 6 (1978), pp. 257–262.

[2] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[3] H. FLANDERS, *An extremal problem on the space of positive definite matrices*, Linear and Multilinear Algebra, 3 (1975), pp. 33–39.

[4] ———, *On the maximal power transfer theorem for n-ports*, Circuit Theory Appl., 4 (1976), pp. 319–344.

[5] E. GILBERT, *Controllability and observability in multivariable control systems*, SIAM J. Control, 1 (1963), pp. 128–151.

[6] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.

[7] R. E. KALMAN, *Mathematical description of linear systems*, SIAM J. Control, 1 (1963), pp. 152–192.

[8] B. C. MOORE, *Principal component analysis in linear systems: controllability, observability and model reduction*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 17–32.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

RECURSIVE LEAST SQUARES ALGORITHM FOR LINEAR PREDICTION PROBLEMS*

SANZHENG QIAO†

Abstract. A new triangularization technique is presented for solving linear prediction problems. The algorithm is based on the exploitation of the special structure that problems of this type exhibit. The reduced triangular system and the error are computed recursively and the problem is solved when the optimal order has been found. The computational complexity of this algorithm is better than existing methods. In addition, good numerical properties are expected of the method.

Key words. linear prediction, Toeplitz, least squares problem, lattice algorithm, orthogonalization, Givens rotation, Householder transformation

AMS(MOS) subject classifications. primary 65F25; secondary 62M20

1. Introduction. Let (t_i) be a time series with a finite number of nonzero terms, say $t_i \neq 0$ for $1 \leqq i \leqq n$. In the linear prediction problem, this signal is modeled by a linear combination of its past values. The predictor parameters $a_1, \cdots, a_p$ $(p < n)$ are then found to minimize the error function

$$E(p) = \sum_{i=1}^{n+p} (t_i + a_1 t_{i-1} + \cdots + a_p t_{i-p})^2. \tag{1.1}$$

This problem can be solved by the autocorrelation normal equations method [9]. Levinson [8] derived an elegant recursive procedure for solving equations of this type. His method is applicable to equations with a general right-hand-side vector. By making use of the special structure possessed by the right-hand-side vector in the normal equations, Durbin [4] derived another method that is twice as fast as Levinson's. However, it is known that the least squares problems are better solved by the orthogonalization methods than by the normal equation techniques [5]. An orthogonalization scheme called the lattice algorithm contributed by Itakura and Saito [6] and Cybenko [2] offers a better alternative. Later Cybenko [3] generalized the lattice algorithm to apply to general Toeplitz matrices. Among the important papers on lattice methods, Makhoul [10] presented a class of stable and efficient algorithms, and Lee, Morf, and Friedlander [7] gave a class of recursive least squares ladder estimation algorithms. Recently, Sweet [13] and Bojanczyk, Brent and de Hoog [1] proposed new QR decomposition algorithms for general Toeplitz matrices. Although the orthogonalization methods have better numerical properties, they usually require more computation.

In this paper, a new recursive algorithm for solving the linear prediction problem is described. This method recursively computes the error $E(p)$ without producing the intermediate solution a_i. When it detects the error curve becoming flat, it then solves the linear prediction problem for the optimal p. This technique is based on Qiao's previous work [11] in which a new fast orthogonal factorization algorithm for general Toeplitz matrices was derived. Numerical experiments showed that this algorithm had good numerical properties. Moreover, by making the use of the special structure, this method is

* Received by the editors September 2, 1986; accepted for publication (in revised form) September 16, 1987. This work was supported by a fellowship funded by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University, Ithaca, New York.

† Center for Applied Mathematics, Cornell University, Ithaca, New York 14853.

almost as fast as the normal equation methods using fast Toeplitz algorithms and five times as fast as the lattice algorithm.

In § 2, the problem setting and some notation is introduced. The algorithm is described in § 3 and a brief remark is made in § 4.

2. Problem setting. The linear prediction problem (1.1) can be formulated as a matrix least-squares problem by writing

$$T(p) = \begin{bmatrix} t_0 & t_{-1} & t_{-2} & \cdots & t_{1-p} \\ t_1 & t_0 & t_{-1} & \cdots & t_{2-p} \\ t_2 & t_1 & t_0 & \cdots & t_{3-p} \\ \vdots & \vdots & \vdots & & \vdots \\ t_{n+p-1} & t_{n+p-2} & t_{n+p-3} & \cdots & t_n \end{bmatrix} \tag{2.1}$$

and

$$a(p) = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix}, \qquad x(p) = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_{n+p} \end{bmatrix}$$

where $t_i = 0$ for $i < 1$ or $i > n$. The problem then becomes

$$\min E(p) = \|T(p)a(p) + x(p)\|_2^2. \tag{2.2}$$

The correlation method, in matrix terms, is the method of normal equations. The solution of the following normal equations:

$$T(p)^T T(p)a(p) = -T(p)^T x(p)$$

is the solution of (2.2). The QR decomposition of $T(p)$,

$$T(p) = Q_1(p)R_1(p) \tag{2.3}$$

where $Q_1(p)$ is an order $n + p$ orthogonal matrix and $R_1(p)$ is $(n + p) \times p$ and an upper triangular matrix, is a better alternative [5], [12]. If $Q_1(p)$ and $R_1(p)$ are partitioned as

$$Q_1(p) = (Q(p), Q_2(p)) \quad \text{and} \quad R_1(p) = \begin{bmatrix} R(p) \\ 0 \end{bmatrix} \tag{2.4}$$

where $Q(p)$ consists of the first p columns of $Q_1(p)$ and $R(p)$ is a $p \times p$ upper triangular matrix, then (2.3) gives

$$T(p) = Q(p)R(p). \tag{2.5}$$

The decomposition above is called an orthogonal factorization (triangularization). Then the solution to (2.2) satisfies the following triangular system:

$$R(p)a(p) = -Q(p)^T x(p). \tag{2.6}$$

In this paper, we assume that $T(p)$ has full column rank, i.e., at least one entry of $T(1)$ is nonzero, and that the diagonal of $R(p)$ is positive. The QR decomposition (2.3) is then unique.

Before we solve (2.2), an important decision that usually has to be made is the determination of an "optimal" p. Clearly, in order to reduce the computation and to minimize the possibility of ill-conditioning, we would like to obtain the minimum value of p which is adequate for the problem (as p increases, $T(p)$ may become more ill-

conditioned). It is known that after some p_0, the error curve remains flat for $p > p_0$ when the signal model is exactly autoregressive [8]. So a simple test to obtain the optimal p is to check when the error curve becomes flat. One suggestion is the use of the following threshold test:

$$1 - \frac{E(p+1)}{E(p)} < \delta. \tag{2.7}$$

When the test above is satisfied for several consecutive values of p, the error curve is then considered to have flattened out. The object of this paper is to present a scheme for the linear prediction problem by computing the orthogonal factorization. While increasing p, this method recursively computes $R(p)$ and applies the test (2.7) without solving for $a(p)$. When the optimal p is reached, it then computes the solution $a(p)$. In terms of computational costs, this method is better than the prevailing methods.

3. The algorithm. Let S denote the circulant shift operator:

$$S = \begin{bmatrix} 0^T & 1 \\ I & 0 \end{bmatrix}.$$

It follows from (2.1) that

$$T(p+1) = (\bar{T}(p), y(p)) = \begin{bmatrix} 0 & 0^T \\ x(p) & T(p) \end{bmatrix} \quad \text{and} \quad x(p+1) = \begin{bmatrix} x(p) \\ 0 \end{bmatrix} \tag{3.1}$$

where

$$\bar{T}(p) = \begin{bmatrix} T(p) \\ 0^T \end{bmatrix} \quad \text{and} \quad y(p) = S^{p+1} x(p+1). \tag{3.2}$$

To derive the algorithm, we need the following lemmas. The proofs can be found in the Appendix.

LEMMA 1. *If $T(p) = Q(p)R(p)$ and $T(p+1) = Q(p+1)R(p+1)$ are two orthogonal factorizations, then $R(p+1)$ and $Q(p+1)$ can be partitioned as*

$$R(p+1) = (\bar{R}(p), r_{p+1}) \quad \text{where } \bar{R}(p) = \begin{bmatrix} R(p) \\ 0^T \end{bmatrix} \tag{3.3a}$$

and

$$Q(p+1) = (\bar{Q}(p), q_{p+1}) \quad \text{where } \bar{Q}(p) = \begin{bmatrix} Q(p) \\ 0^T \end{bmatrix}. \tag{3.3b}$$

LEMMA 2. *Define*

$$u(p) = Q(p)^T x(p); \tag{3.4}$$

then

$$u(p+1) = \begin{bmatrix} u(p) \\ u_{p+1} \end{bmatrix} \tag{3.5}$$

and

$$E(p+1) = E(p) - u_{p+1}^2. \tag{3.6}$$

Lemma 1 says the first p columns of $R(p+1)(Q(p+1))$ are identical to the columns of $R(p)(Q(p))$ except for the zeros at the bottom. Lemma 2 indicates that $u(p)$ is a

subvector of $u(p+1)$ and $E(p+1)$ can be obtained by subtracting u_p^2 from $E(p)$. Then it is clear that $R(p)$, $u(p)$ and $E(p)$ are given recursively, if u_p and r_{p+1} can be computed from $u(p-1)$ and r_p.

We first describe how u_p is given by r_p and $u(p-1)$. The definition of $u(p)$ in (3.4) implies that $u(p)$ is the solution of the triangular system

$$R(p)^T u(p) = T(p)^T x(p).$$

Then by Lemma 1, u_p is given by the last equation of the above system, and using (3.1) and (3.2) we get

$$r_p^T u(p) = (S^p x(p))^T x(p). \tag{3.7}$$

Second, we show the computation of r_{p+1}. Suppose $T(p)$ has QR decomposition (2.3), and that from (3.1) we have

$$\begin{bmatrix} 0 & Q_1(p)^T \\ 1 & 0^T \end{bmatrix} T(p+1) = \begin{bmatrix} u(p) & R(p) \\ v(p) & 0 \\ 0 & 0^T \end{bmatrix} \tag{3.8}$$

where vector $v(p) = Q_2(p)^T x(p)$. Then the matrix in (3.8) is triangularized by a product of Givens rotations $W = W_1 \cdots W_{n+p}$, i.e.,

$$W_1 \cdots W_{n+p} \begin{bmatrix} u(p) \\ v(p) \\ 0 \end{bmatrix} = \beta e_1$$

where each W_i is an $(i, i+1)$-plane rotation and $e_1 = (1, 0, \cdots, 0)^T$. If we define

$$\eta_{i+1}^2 = \eta_i^2 - u_i^2 \quad \text{for } 1 \leqq i \leqq p \quad \text{with } \eta_1^2 = \|x(p)\|_2^2, \tag{3.9}$$

then each W_i $(i = 1, \cdots, p)$ is determined by the following equation:

$$\begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix} \begin{bmatrix} u_i \\ \eta_{i+1} \end{bmatrix} = \begin{bmatrix} \eta_i \\ 0 \end{bmatrix}$$

where c_i and s_i are the cosine and sine parameters for the rotation W_i. Specifically,

$$c_i = u_i/\eta_i, \qquad s_i = \eta_{i+1}/\eta_i. \tag{3.10}$$

Assume W is chosen so that the diagonal of the resulting triangular matrix is positive. Because of (3.5), the quantities η_p, $W_1, \cdots, W_{p-1}$ are the same as for the lower-order systems. We can therefore assume that η_p, $W_1, \cdots, W_{p-1}$ are calculated in the previous steps. By the above argument, η_{p+1} and W_p are given by (3.9) and (3.10) with $i = p$. Now the desired r_{p+1} is obtained by

$$W_1 \cdots W_p \begin{bmatrix} r_p \\ 0 \end{bmatrix}. \tag{3.11}$$

Noting that r_p is not affected by $W_{p+1}, \cdots, W_{n+p}$ and that QR decomposition is unique, we get

$$W_1 \cdots W_{n+p} \begin{bmatrix} u(p) & R(p) \\ v(p) & 0 \\ 0 & 0^T \end{bmatrix} = R(p+1).$$

Finally, by comparing (3.9) with (3.6), we see that

$$\eta_{p+1}^2 = E(p) \quad \text{with } E(0) = \|x(p)\|_2^2.$$

Thus the test (2.7) is simply

$$1-\frac{E(p+1)}{E(p)}=1-\frac{\eta_{p+2}^2}{\eta_{p+1}^2}=1-s_{p+1}^2=c_{p+1}^2<\delta. \tag{3.12}$$

We conclude this section by presenting the following algorithm.

ALGORITHM.
Initialize $p = 1$, $r_p = \|x(p)\|_2$ and $\eta_p^2 = \|x(p)\|_2^2$,
repeat
Solve u_p from $u(p-1)$ and r_p using (3.7);
Update $\eta_{p+1}^2 = \eta_p^2 - u_p^2$,
Compute W_p from (3.10);
Find r_{p+1} by (3.11);
$p \leftarrow p + 1$;
Until $c_{p-1}^2 < \delta$ for several consecutive steps;
Solve u_p from r_p and $u(p-1)$ using (3.7);
Back solve $R(p)a(p) = -u(p)$.

If we refer to a floating point multiplication and addition as a FLOP, then the algorithm above requires only about $np + 3p^2/2$ FLOPS. The lattice algorithm needs approximately $5np$ FLOPS [2].

4. Concluding remarks. This paper has described a triangularization approach to solving the linear prediction problem. This algorithm is among the most efficient methods yet devised. Moreover, it computes the error $E(p)$ recursively and finds the optimal p. A slight modification can adapt the method to solving the discrete-time Wiener filtering problem, in which t_i in (1.1) is replaced by x_i so that the right-hand-side vector $x(p)$ is a general column vector.

Appendix. Proofs of lemmas.

Proof of Lemma 1. Let $Q(p+1)$ and $R(p+1)$ be partitioned as follows:

$$Q(p+1)=[\bar{Q}'(p), q_{p+1}] \quad \text{and} \quad R(p+1)=[\bar{R}'(p), r_{p+1}]$$

where

$$\bar{Q}'(p)=\begin{bmatrix} Q'(p) \\ h^T \end{bmatrix} \quad \text{and} \quad \bar{R}'(p)=\begin{bmatrix} R'(p) \\ 0^T \end{bmatrix}.$$

By the partitioning of $T(p+1)$ given in (3.1) and (3.2), we get

$$T(p)=Q'(p)R'(p) \quad \text{and} \quad 0^T=h^TR'(p). \tag{A1}$$

So, because $R'(p)$ is nonsingular, the vector

$$h=0 \tag{A2}$$

is null, implying that $Q'(p)$ must be orthogonal, for otherwise the first p columns of $Q(p+1)$ are not orthogonal. Thus, the first equation in (A1) is the orthogonal decomposition of $T(p)$. Hence $Q'(p) = Q(p)$ and $R'(p) = R(p)$, under the assumption that diagonal elements of both $R'(p)$ and $R(p)$ are positive. These relations, together with (A2), complete the proof.

Proof of Lemma 2. The definition of $u(p)$ implies that

$$R(p+1)^T u(p+1) = T(p+1)^T x(p+1). \tag{A3}$$

Denote

$$u(p+1) = \begin{bmatrix} z \\ u_{p+1} \end{bmatrix}. \tag{A4}$$

It then suffices to show that the vector z satisfies

$$R(p)^T z = T(p)^T x(p),$$

which can be done by partitioning (A3) and substituting (3.3a), (A4), (3.1), and (3.2).

The second part of the lemma follows from the first part, (3.1), and

$$E(p) = \|x(p)\|_2^2 - \|u(p)\|_2^2.$$

This can be proved by noting (2.3), (2.4), (2.6) and

$$\begin{aligned} E(p) &= \|T(p)a(p) + x(p)\|_2^2 \\ &= \|Q_1(p)(T(p)a(p) + x(p))\|_2^2 \\ &= \|R(p)a(p) + Q(p)^T x(p)\|_2^2 + \|Q_2(p)^T x(p)\|_2^2 \\ &= \|Q_2(p)^T x(p)\|_2^2 \\ &= \|x(p)\|_2^2 - \|Q(p)^T x(p)\|_2^2. \end{aligned}$$

Acknowledgment. The author thanks David Schimmel for his thorough reading of the draft of this paper, F. Luk for his many helpful comments, and an anonymous referee for valuable suggestions and a clearer proof of Lemma 1.

REFERENCES

[1] A. W. Bojanczyk, R. P. Brent, and F. R. de Hoog, *QR factorization of Toeplitz matrices*, Numer. Math., 49 (1986), pp. 81–94.
[2] G. Cybenko, *A general orthogonalization technique with applications to time series analysis and signal processing*, Math. Comput., 40 (1983), pp. 323–336.
[3] ———, *Fast Toeplitz orthogonalization using inner products*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 734–740.
[4] J. Durbin, *The fitting of time-series models*, Rev. Inst. Internat. Statist., 28 (1960), pp. 233–243.
[5] G. H. Golub, *Numerical methods for solving linear least squares problems*, Numer. Math., 7 (1965), pp. 206–216.
[6] F. Itakura and S. Saito, *Digital filtering techniques for speech analysis and synthesis*, Proc. 7th Internat. Congress on Acoustics, Budapest, 1971, pp. 261–264.
[7] D. T. L. Lee, M. Morf, and B. Friedlander, *Recursive least squares ladder estimation algorithms*, IEEE Trans. Acoust. Speech Signal Process., ASSP-29 (1981), pp. 627–641.
[8] N. Levinson, *The Wiener* rms *(root-mean-square) error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.
[9] J. Makhoul, *Linear prediction: a tutorial review*, Proc. IEEE., 63 (1975), pp. 561–580.
[10] ———, *Stable and efficient lattice methods for linear prediction*, IEEE Trans. Acoust., Speech Signal Process., ASSP-25(1977), pp. 423–428.
[11] S. Qiao, *A hybrid algorithm for fast Toeplitz orthogonalization*, Numer. Math., to appear.
[12] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[13] D. R. Sweet, *Fast Toeplitz orthogonalization*, Numer. Math., 43 (1984), pp. 1–21.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

004

A GUIDE TO THE ACCELERATION OF ITERATIVE METHODS WHOSE ITERATION MATRIX IS NONNEGATIVE AND CONVERGENT*

G. AVDELAS†, J. DE PILLIS‡, A. HADJIDIMOS†§, AND M. NEUMANN§

Abstract. For an $n \times n$ nonnegative matrix B whose spectral radius is less than unity we consider the acceleration of the fixed-point iteration scheme

$$x_{j+1} = Bx_j + c \tag{1}$$

by two parameter-dependent techniques: the extrapolation method

$$z_{j+1} = [\omega B + (1-\omega)B]z_j + \omega c \tag{2}$$

and the second-degree stationary method

$$u_{j+1} = \omega B u_j + (1-\omega)u_{j-1} + \omega c. \tag{3}$$

It is shown whether, when B is (also) irreducible, it is possible to accelerate (1) and, if so, whether technique (2) or (3) provides the best acceleration, which is largely determined by the cyclicity p of B. In this paper all possible values of p are analyzed. In the case when B is reducible, the possibilities for accelerating (1) can be determined with the aid of the Frobenius normal form of B.

Actually, one motivation for the present work is an observation that if B is an $n \times n$ nonnegative matrix whose spectral radius is less than 1, then for *no* decomposition of B into $B = B_1 + B_2$, where both B_1 and B_2 are nonnegative matrices, does the second-degree iterative method

$$w_{j+1} = B_1 w_j + B_2 w_{j-1} + c$$

attain a convergence rate superior to the scheme in (1).

Key words. iterative methods, acceleration of convergence, nonnegative matrices

AMS(MOS) subject classifications. primary 65F10; secondary 15A06, 15A18

1. Introduction.

In numerical linear algebra, fixed-point iteration, schemes of the form

$$x_{j+1} = Bx_j + c, \qquad j = 0, 1, \cdots \tag{1.1}$$

are often used to compute an approximate solution to the nonsingular system

$$Ax = (I - B)x = c. \tag{1.2}$$

It is well known that the iteration scheme (1.1) converges to the solution to (1.2) from every initial vector x_0 if and only if $\rho(B)$ (here $\rho(\cdot)$ denotes the *spectral radius* of a matrix) is less than unity. Moreover, it is customary to adopt the quantity $\rho(B)$ (or sometimes $-\ln(\rho(B))$) as *a measure* of the rate of convergence of (1.1) to the solution to (1.2).

If $\rho(B)$ is not much less than 1, then the scheme (1.1) has a poor rate of convergence and we may seek to accelerate it. Of the several techniques which have been suggested for this purpose in the literature, two will be of interest to us here. The first technique consists of converting scheme (1.1) to a *second degree stationary linear iteration process*

* Received by the editors July 13, 1987; accepted for publication (in revised form) October 13, 1987.
† General Department, Technical University of Crete, Hania, Greece.
‡ Department of Mathematics, University of California, Riverside, California 92521.
§ Department of Mathematics, University of Connecticut, Storrs, Connecticut 06268. The research of this author was supported in part by National Science Foundation grant DMS-8700604.

which has the form

$$y_{j+1} = B_1 y_j + B_2 y_{j-1} + c, \qquad j = 0, 1, \cdots . \tag{1.3}$$

The second technique consists of introducing an iteration parameter $\omega \in \mathbb{R}\backslash\{0\}$, frequently referred to as an *extrapolation parameter*, and performing the *extrapolated* iteration process

$$z_{j+1} = B_\omega z_j + \omega c, \qquad j = 0, 1, \cdots , \tag{1.4}$$

where

$$B_\omega := \omega B + (1 - \omega) I. \tag{1.5}$$

We then ask for an $\omega = \omega_{\text{opt}}$ which minimizes $\rho(B_\omega)$ (see, for example, Isaacson and Keller [1966, pp. 73–78]).

Consider for a moment the second-degree scheme of (1.3). It is readily observed that this scheme can be embedded, in fact identified with, the $2n$-dimensional first-degree iterative scheme given by

$$\begin{bmatrix} y_{j+1} \\ y_j \end{bmatrix} = \begin{bmatrix} B_1 & B_2 \\ I & 0 \end{bmatrix} \begin{bmatrix} y_j \\ y_{j-1} \end{bmatrix} + \begin{bmatrix} c \\ 0 \end{bmatrix}, \tag{1.6}$$

showing that we can use

$$\rho\left(\begin{bmatrix} B_1 & B_2 \\ I & 0 \end{bmatrix}\right) \tag{1.7}$$

to measure the rate of convergence of (1.3).

In this paper we shall only be concerned with the acceleration of fixed-point iteration schemes (1.1) in which the matrix $B = (b_{ij})$ is nonnegative, that is, $b_{ij} \geqq 0$, $i, j = 1, \cdots , n$. We shall always assume that $\rho(B) < 1$. Then, as is well known, $A = I - B$ is a nonsingular M-matrix. We mention that many of the properties of nonsingular M-matrices, and the various applications which give rise to these matrices, are discussed in Berman and Plemmons [1979] and Varga [1962].

In the first result of this paper (cf. Theorem 2.1) we show that if the second-degree scheme (1.3) is obtained from (1.1) by letting

$$B = B_1 + B_2, \qquad B_1, B_2 \gneqq 0, \tag{1.8}$$

then (1.3) *does not* possess a superior rate of convergence to (1.1). This result shows that for $0 < \omega \leqq 1$, the second-degree stationary iteration method

$$u_{j+1} = \omega B u_j + (1 - \omega) u_{j-1} + \omega c, \tag{1.9}$$

which we derive from (1.4) by setting $B_1 = \omega B$ and $B_2 = (1 - \omega) I$, *does not* possess a better rate of convergence than the scheme in (1.4) for ω's in this range. We are therefore led to ask the following questions:

(a) Under what additional assumptions on B does the second degree iteration scheme in (1.9) have, for some $\omega \in R\backslash[0, 1]$, a superior rate of convergence to the extrapolation scheme (1.4) when the extrapolation parameter varies over all ω's for which (1.4) converges and vice versa?

(b) Under what additional conditions on B do neither of the schemes (1.4) and (1.9) provide a means to accelerate the convergence of the "original" scheme (1.1)?

We shall attack the above questions by assuming at first that our matrix B, which we have previously assumed to be nonnegative with $\rho(B) < 1$, is also irreducible. We

shall then use the Frobenius normal form of a matrix to consider the answer to the same questions when B is reducible. As background information concerning the question in (a), we mention that in the case when B is irreducible it can be shown, from the papers by de Pillis and Neumann [1981], Hadjidimos [1983], [1986], Hughes-Hallett [1981], and Kulisch [1968], that

$$\rho(B_\omega)<1 \quad \text{when} \quad \begin{cases} 0<\omega<\dfrac{2}{1+\rho(B)}=:\omega' & \text{if } B \text{ is } p\text{-cyclic}, \quad p\geqq 2, \qquad (1.10a) \\ \omega\in(0,\omega'') \quad \text{for some } \omega''>\omega' & \text{if } B \text{ is primitive.} \qquad (1.10b) \end{cases}$$

In Lemma 2.1 we show that when B is irreducible, the range of ω's for which the second-degree scheme in (1.9) converges is at least ω' given in (1.10a).

The answers which we provide to the above questions can be briefly summarized as follows:

(i) When B is irreducible p-cyclic, $p \geqq 3$, then neither scheme (1.4) nor scheme (1.9) provides for any ω a rate of convergence which is superior to (1.1).

(ii) When B is irreducible and 2-cyclic, then the second-degree iteration scheme of (1.9) attains for some $\omega = \hat{\omega}_{\text{opt}} > 1$ a rate of convergence which is superior to both the "original" scheme (1.1) and the extrapolation scheme (1.4).

(iii) When B is irreducible and primitive, then both schemes (1.4) and (1.9) attain a rate of convergence superior to (1.1). However, in general, which of the acceleration schemes (1.4) or (1.9) will provide optimal acceleration to (1.1) can depend on the structure of the convex hull of the eigenvalues of B.

The results above are stated rigorously in Theorem 3.1 and we use several lemmas preceding this theorem to provide a substantial part of its proof. The discussion on the acceleration of the scheme in (1.1) in case B is reducible is given following that theorem.

We finally mention that most of the proofs to our results in § 3 are based on the construction of capturing ellipses and the determination of such ellipses which are optimal. In so doing we utilized results of de Pillis [1980], Avdelas, Galanis, and Hadjidimos [1983], Young [1971, Chap. 6], and Young and Eidson [1970].

2. Further background material and initial results. Let $B = (b_{ij})$ be an $n \times n$ real or complex matrix. B is called *irreducible* if for *no* permutation matrix P,

$$PBP^T = \begin{bmatrix} \tilde{B}_{11} & \tilde{B}_{12} \\ 0 & \tilde{B}_{22} \end{bmatrix}, \qquad (2.1)$$

where $\tilde{B}_{11}$ and $\tilde{B}_{22}$ are square matrices. It is well known that the irreducibility of B is equivalent to its directed graph being strongly connected. If for some permutation matrix P, (2.1) holds, then B is said to be *reducible*. In this case there exists a permutation matrix Q such that

$$QBQ^T = \begin{bmatrix} \tilde{B}_{11} & \cdots & \tilde{B}_{1k} \\ & \tilde{B}_{22} & \\ & & \ddots \\ 0 & & \tilde{B}_{kk} \end{bmatrix}, \qquad (2.2)$$

where each diagonal block is square and irreducible or the 1×1 null matrix. The matrix on the right-hand side of (2.2) is called the *Frobenius normal form of* B (e.g., Berman and Plemmons [1979], Varga [1962]).

Suppose now that B is an $n \times n$ nonnegative and irreducible matrix. Then the spectral radius of B is a positive eigenvalue of B. Moreover each eigenvalue λ of B with $|\lambda| = \rho(B)$ is simple. The number of distinct eigenvalues of B with moduli $\rho(B)$ is known as the *cyclicity* of B. If the cyclicity of B is p, then the eigenvalues λ of B with $|\lambda| = \rho(B)$ are given by $\lambda_j = \rho(B)e^{2\pi j/p}$, $j = 0, 1, \cdots, p-1$. When $p = 1$, B is called a *primitive* matrix. If $p > 1$, then according to a theorem of Frobenius (see Varga [1962, Thm. 2.8]) there exists a permutation matrix P such that

$$PBP^T = \begin{bmatrix} 0 & \tilde{B}_{12} & & \\ & \ddots & \ddots & \\ & & \ddots & \tilde{B}_{p-1,p} \\ \tilde{B}_{p1} & & & 0 \end{bmatrix},$$

where the diagonal blocks are all null square matrices.

In the first result of this paper we shall show that the second-degree stationary iterative scheme (1.3), induced by the iterative process (1.1) in accordance with (1.8), *does not* lead to a scheme with a superior rate of convergence compared to (1.1).

THEOREM 2.1. *Let B be an $n \times n$ nonnegative matrix with $\rho(B) < 1$. Consider the iterative schemes* (1.1) *and* (1.4), *where B_1 and B_2 satisfy* (1.8). *Then*

$$\rho(B) \leqq \rho\left(\begin{bmatrix} B_1 & B_2 \\ I & 0 \end{bmatrix}\right) \leqq \rho^{1/2}(B). \tag{2.3}$$

Moreover, when B is irreducible, then the inequalities in (2.3) *are all strict.*

Proof. Recall that a splitting of a square matrix C into $C = M - N$ is called regular if $\det(M) \neq 0$, $M^{-1} \geqq 0$, and $N \geqq 0$. In our proof we shall make use of a *comparison theorem* for the rate of convergence of iteration matrices induced by regular splittings due to Varga [1962, Thm. 3.15].

Consider the three splittings of the $(2n) \times (2n)$ matrix C given by

$$C := \begin{bmatrix} I & -B \\ -I & I \end{bmatrix} \tag{2.4}$$

$$= \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} 0 & B \\ I & 0 \end{bmatrix} \tag{2.5a}$$

$$= \begin{bmatrix} I & -B_1 \\ 0 & I \end{bmatrix} - \begin{bmatrix} 0 & B_2 \\ I & 0 \end{bmatrix} \tag{2.5b}$$

$$= \begin{bmatrix} I & -B \\ 0 & I \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix}. \tag{2.5c}$$

All three splittings are easily observed to be regular. Moreover,

$$\begin{bmatrix} 0 & B \\ I & 0 \end{bmatrix} \geqq \begin{bmatrix} 0 & B_2 \\ I & 0 \end{bmatrix} \geqq \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix}. \tag{2.6}$$

Now, as $\rho(B) < 1$, C is invertible and can be expanded in a Neumann series. This easily shows that C has a nonnegative inverse. Hence, by Theorem 3.13 in Varga [1962], all three iteration matrices induced by the splittings in (2.5a–c) have a spectral radius less

than 1. Moreover by Varga's comparison theorem mentioned above we have that

$$\rho(B) = \rho\left\{\begin{bmatrix} I & B \\ 0 & I \end{bmatrix}\begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix}\right\}$$
$$\leqq \rho\left\{\begin{bmatrix} I & B_1 \\ 0 & I \end{bmatrix}\begin{bmatrix} 0 & B_2 \\ I & 0 \end{bmatrix}\right\}$$
$$\leqq \rho\left\{\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} 0 & B \\ I & 0 \end{bmatrix}\right\} = \rho^{1/2}(B),$$

and so (2.3) is valid. To conclude that strict inequalities hold in (2.3) when B is irreducible, we need only observe that since $A = I - B$ is now a nonsingular and irreducible M-matrix, $A^{-1} > 0$, showing that

$$C^{-1} = \begin{bmatrix} A^{-1} & A^{-1}B \\ A^{-1} & I + A^{-1}B \end{bmatrix} > 0$$

and the result now follows by Corollary 2 to Theorem 3.15 in Varga [1962]. □

An immediate consequence of our theorem is the following statement, the proof of which is omitted.

COROLLARY 2.1. *If B is an $n \times n$ nonnegative matrix with $\rho(B) < 1$, then for no $\omega \in (0, 1]$ is the rate of convergence of the iterative scheme* (1.9) *superior to the rate of convergence of the iterative scheme* (1.4).

In view of Corollary 2.1, (1.10), and Theorem 2.2 below, it will be seen that to obtain a complete comparison between schemes (1.4) and (1.9) we need only consider their behaviour in the case where $\omega > 1$.

We have already mentioned in § 1 that (1.10) gives an interval of convergence in ω for the extrapolation scheme (1.4) when B is an $n \times n$ nonnegative and irreducible matrix with $\rho(B) < 1$. In much of the remainder of this section we summarize results from the literature concerning the convergence of the second degree scheme in (1.9) which will be needed in the development of our work in the next section. To this end we need first to introduce some further notation and terminology.

Let B be an $n \times n$ matrix. We shall denote by SD (B) its *spectral disc*, that is,

$$\text{SD}(B) = \{z \in \mathbb{C} \mid |z| \leqq \rho(B)\}.$$

The spectral circle of B is then SC (B) = ∂(SD (B)), $\partial(\cdot)$ denoting the *boundary* of a set. CH (B) will denote the *convex hull* of the spectrum of B, that is,

$$\text{CH}(B) = \{z \in \mathbb{C} \mid z = \Sigma_k \alpha_k \lambda_k \text{ with } \alpha_k \geqq 0, \Sigma_k \alpha_k = 1, \text{ and } \lambda_k \in \sigma(B)\}.$$

$\hat{\text{CH}}$ (B) will denote the *smallest convex polygon which is symmetric with respect to the axes*. Thus if S is the infinite strip

$$S := \{z \mid -1 < \text{Re}(z) < 1\},$$

then

$$\sigma(B) \subset \text{CH}(B) \subset \hat{\text{CH}}(B) \subset \text{SD}(B) \subset S. \tag{2.7}$$

We next define the notion of capturing ellipses which is essentially taken from de Pillis [1980] and Young [1971, Chap. 6]. Beforehand let us comment that, since all the ellipses which we shall consider lie in the complex plane and are symmetric about the real and imaginary axes, it will be convenient to refer to their appropriate semi-axis as either the *real semi-axis* or the *imaginary semi-axis*, as the case may be.

DEFINITION 2.1. Let K be a subset of S. Then a capturing *ellipse for* K is an ellipse E_K that is symmetric about the axes, passes through at least one point in K, contains K, and lies in S. In particular, if B is an $n \times n$ matrix such that $\sigma(B) \subseteqq S$, then a *capturing ellipse* is a capturing ellipse for $\sigma(B)$.

If E is an ellipse which is symmetric about the axes and contained in S, it will be convenient to denote the length of its real semi-axis by $M_r(E)$ and length of its imaginary semi-axis by $M_i(E)$. As a *measure of the eccentricity* of E we shall adopt the ratio

$$\lambda_E = \frac{M_r(E) - M_i(E)}{M_r(E) + M_i(E)}. \tag{2.8}$$

Note that because $E \subset S$, the quantity

$$\mu_E := \frac{M_r(E) + M_i(E)}{1 + [1 - M_r^2(E) + M_i^2(E)]^{1/2}} \tag{2.9}$$

is well defined and strictly bounded above by 1. The importance of the concept of a capturing ellipse is illustrated by the following theorem.

THEOREM 2.2. (de Pillis [1980], Avdelas, Galanis, and Hadjidimos [1983]). *Let B be an $n \times n$ matrix whose spectrum lies in S and let E be a capturing ellipse. Then for*

$$\omega = \omega(E) = 1 + \lambda_E \mu_E^2, \tag{2.10}$$

$\omega > 0$ and the iterative scheme (1.9) *converges to the solution to* (1.2) *from any arbitrary vectors $u_{-1}, u_0 \in R^n$. Moreover,*

$$\rho\left(\begin{bmatrix} \omega B & (1-\omega)I \\ I & 0 \end{bmatrix}\right) = \mu_E. \tag{2.11}$$

Conversely, for any $0 < \omega < 2$ for which the scheme (1.9) *converges with a convergence rate*

$$\rho\left(\begin{bmatrix} \omega B & (1-\omega)I \\ I & 0 \end{bmatrix}\right) =: \mu,$$

there exists a unique capturing ellipse E such that $\mu_E = \mu$, whose eccentricity is given by $\lambda(E) = (\omega - 1)/\mu^2$, whose real semi-axis is given by

$$M_r(E) = \begin{cases} \mu, & \omega = 1, \\ \dfrac{\mu^2 + (\omega - 1)}{\omega\mu}, & \omega \neq 1, \end{cases}$$

and whose imaginary semi-axis is given by

$$M_i(E) = \begin{cases} \mu, & \omega = 1, \\ \dfrac{\mu^2 - (\omega - 1)}{\omega\mu}, & \omega \neq 1. \end{cases}$$

Theorem 2.2, and in particular its result in (2.11), motivate our next definition.

DEFINITION 2.2. (i) Let P be a point in S and let $\mathscr{F}_P$ be the family of all capturing ellipses through P. An *optimal capturing ellipse* for P is the capturing ellipse in $\mathscr{F}_P$ which minimizes μ_{E_P} as E_P varies over $\mathscr{F}_P$.

(ii) Let K be a subset of S and let $\mathscr{F}_K$ be the family of all capturing ellipses for K. An *optimal capturing ellipse* for K is the capturing ellipse in $\mathscr{F}_K$ which minimizes μ_{E_K} as E_K varies over $\mathscr{F}_K$.

In the next section as a general rule we shall use the "hat" symbol to denote capturing ellipses which are optimal, e.g., for a point $P \in S$, $\hat{E}_P$ denotes the optimal capturing ellipse through P. Moreover, under the conditions of Theorem 2.2 on B, we shall denote by $\hat{\omega}_{\text{opt}}$ the value of ω given in (2.10) which is determined when setting $E = \hat{E}$ in (2.8) and (2.9). Thus according to the equivalent statements of the theorem and according to the discussion in § 1, $\hat{\omega}_{\text{opt}}$ furnishes the iteration parameter for which the second-degree scheme of (1.9) attains the optimal rate of convergence.

Next, using Theorem 2.2 we are now able to show that if B is an $n \times n$ nonnegative and irreducible matrix with $\rho(B) < 1$, then the interval of convergence in ω of the second-degree stationary iterative scheme of (1.9) always extends to at least ω' given in (1.10a). Thus, by (1.10a), when B is also p-cyclic with $p \geqq 2$, then the second-degree scheme always possesses an interval of convergence in ω which is at least as large as the interval of convergence in ω of the extrapolation scheme (1.4).

LEMMA 2.1. *Let B be an $n \times n$ nonnegative and irreducible matrix with $\rho(B) < 1$. Then the interval of convergence in ω, $(0, \tilde{\omega})$, of the second-degree scheme* (1.9) *satisfies that $\tilde{\omega} > \omega'$, where ω' is as given in* (1.10a).

Proof. According to Theorem 2.1 the interval of convergence of (1.9) in ω is at least $(0, 1]$. Consider now the family $\mathcal{F}$ of all ellipses E possessing the following properties:

(i) Each $E \in \mathcal{F}$ is symmetric with respect to the axes and has a real semi-axis $M_r(E) = 1$.

(ii) $\sigma(B) \subset E$ for all $E \in \mathcal{F}$.

For each $E \in \mathcal{F}$ let $\tilde{E}$ denote its union with its interior. Let

$$\mathcal{E} = \partial\left(\bigcap_{E \in \mathcal{F}} \tilde{E}\right).$$

Then $\mathcal{E}$ is an ellipse with $M_r(\mathcal{E}) = 1$ which contains $\sigma(B)$. Evidently, $\mathcal{E}$ has $\rho(B) \geqq M_i(\mathcal{E})$ with equality if and only if $i\rho(B) \in \sigma(B)$. It is now clear by a limiting argument involving (2.8)–(2.10) that, by Theorem 2.2, the interval of convergence in ω of (1.9) extends from $(0, 1]$ to $(0, \tilde{\omega})$, where

$$\tilde{\omega} = \frac{2}{1 + M_i(\mathcal{E})} \geqq \omega'. \qquad \square$$

Throughout most of the next section we shall continue to assume that B is an $n \times n$ nonnegative matrix with $\rho(B) < 1$. We shall analyse and compare the rates of convergence of the extrapolation scheme (1.4), when ω satisfies (1.10) with the rate of convergence of (1.9) for all ω's for which it converges.

3. Comparison between the extrapolated and the second degree schemes. Our comparison of the iterative schemes given in (1.4) and (1.9) will concentrate on determining the value of ω for which each of the schemes achieves an optimal convergence rate. It will be seen that in several cases the optimal value of ω for either scheme is dependent only on the cyclicity p of the matrix B.

We begin with the following observation.

LEMMA 3.1. *Let B be an $n \times n$ irreducible nonnegative with $\rho(B) < 1$ and cyclicity $p \geqq 2$. Then for the extrapolated scheme* (1.4), $\omega_{\text{opt}} = 1$.

Proof. We appeal here to Theorem 3.2 in de Pillis and Neumann [1981]. Since B is irreducible of cyclicity $p \geqq 2$, B has p eigenvalues uniformally distributed on SC (B). But then SC (B) satisfies the requirements for being a capturing circle given in axioms (3.5) of de Pillis and Neumann [1981]. Since SC (B) is centered at the origin, the conclusion now follows by Theorem 3.2 cited above. $\square$

The above result says, of course, that for $p \geqq 2$ the optimal extrapolation parameter for (1.4) is independent of p. We now proceed to show that the optimal iteration parameter $\hat{\omega}_{\text{opt}}$ for (1.9) is a constant equal to 1 only when $p \geqq 3$. Our analysis here utilizes the notion of capturing ellipses introduced in the previous section. It is motivated by results in Avdelas, Galanis, and Hadjidimos [1983] which, in turn, are based on an algorithm of Young and Eidson [1970]. Because of the symmetry about the axes of the capturing ellipses, Young and Eidson conclude that the optimal capturing ellipse for $\sigma(B)$ coincides with the optimal capturing ellipse of $C\hat{H}(B)$. They show that the latter can be found by considering, in a certain order, optimal capturing ellipses through vertices of $C\hat{H}(B)$ in the first quadrant only.

In our analysis here we shall make use of the following simple observation.

LEMMA 3.2. *Let $P(\alpha, \beta) \in S\backslash\{(0, 0)\}$ be a point in the first quadrant and let $\hat{E}_P$ be the optimal capturing ellipse through P. A sufficient condition for the eccentricity $\lambda_{\hat{E}_P} < 0$ is that $\alpha \leqq \beta$.*

Proof. According to a simplified form of the equations in (4.24) in Young [1971, p. 196] given in Avdelas, Galanis, and Hadjidimos [1983], $\mu_{\hat{E}_P} > 0$,

$$M_r(\hat{E}_P) = [2\mu_{\hat{E}_P}\alpha^2/(1 + \mu_{\hat{E}_P}^2)]^{1/3}$$

and

$$M_i(\hat{E}_P) = [2\mu_{\hat{E}_P}\beta^2/(1 - \mu_{\hat{E}_P}^2)]^{1/3}.$$

Hence $M_r(\hat{E}_P) < M_i(\hat{E}_P)$. □

We are now in a position to prove the following lemma.

LEMMA 3.3. *Let B be an $n \times n$ nonnegative and irreducible matrix of cyclicity $p \geqq 3$. Then the optimal iteration parameter for* (1.9) *is* $\hat{\omega}_{\text{opt}} = 1$.

Proof. Consider the eigenvalues of B on SC (B) and construct the smallest regular q-gon G, symmetric with respect to both axes, such that all these eigenvalues are vertices of G. Observe that if p is even then $q = p$, while if p is odd then $q = 2p$. Certainly $G \subseteq C\hat{H}(B)$ (see (2.7)) and all vertices of G are vertices of $C\hat{H}(B)$.

We distinguish between two cases: $q(\text{mod } 4) = 0$ and $q(\text{mod } 4) = 2$.

Case 1. $q(\text{mod } 4) = 0$. In this case, as illustrated in Fig. 1, G, and hence $C\hat{H}(B)$, have vertices on both axes, $P(\rho(B), 0)$ and $Q(0, \rho(B))$, respectively. Then the optimal capturing ellipse $\hat{E}$ for $C\hat{H}(B)$, and hence for $\sigma(B)$, must have semi-axes $M_r(\hat{E}), M_i(\hat{E}) \geqq \rho(B)$. But then according to the results given in displays (4.17)–(4.18) of Young [1971, Chap. 6], necessarily, $M_r(E) = M_i(E) = \rho(B)$ showing that $\hat{\omega}_{\text{opt}} = 1$ by (2.8)–(2.10).

Case 2. $q(\text{mod } 4) = 2$. In this case, as illustrated in Fig. 2, G has a horizontal edge linking its highest point in the first quadrant, $Q := Q(\rho(B) \cos\theta, \rho(B) \sin\theta)$, where $\theta = (\frac{1}{2} - 1/q)\pi$, with its adjacent vertex T in the second quadrant. Moreover, the ordinate of Q is always greater than its abscissa. To determine the optimal capturing ellipse for $C\hat{H}(B)$ we now follow the algorithm in Young and Eidson [1970] which is implicitly given in Young [1971, Chap. 6].

We begin by observing that any capturing ellipse for $C\hat{H}(B)$, and hence for $\sigma(B)$, must contain both $P(\rho(B), 0)$ and Q. It must therefore contain (also) all vertices of $C\hat{H}(B)$ other than P and Q which lie in the region bounded by the spectral circle and the chord joining P and Q (see Fig. 2). Thus in applying the Young–Eidson [1970] algorithm for determining the optimal capturing ellipse for $C\hat{H}(B)$ all such vertices can be discarded.

Consider the ellipse Σ symmetric about the axes which passes through the points $(1, 0)$, Q, T, and $(-1, 0)$ and note that any capturing ellipse for Q or for a set which

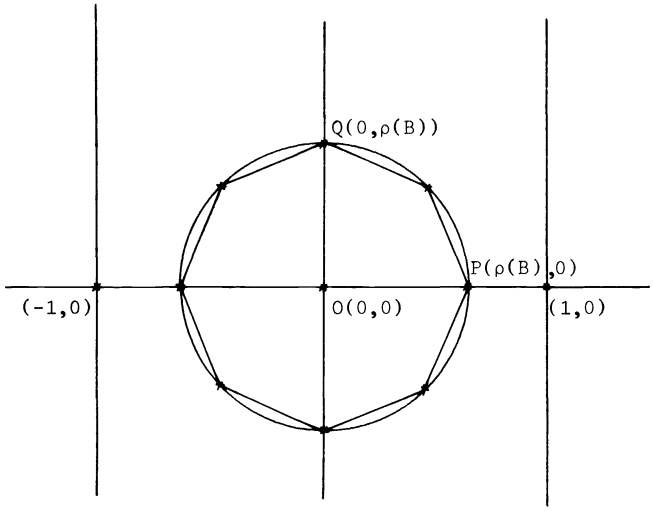


FIG. 1

contains Q must (also) contain, because of its symmetry about the axes, all vertices of $C\hat{H}(B)$ in the first quadrant other than Q which lies in the region bounded by Σ and the line segment joining Q and T. Once again the Young–Eidson algorithm permits us to discard all such vertices in determining the optimal capturing ellipse for $C\hat{H}(B)$. Thus to complete the construction of the optimal ellipse it remains to consider all vertices of $C\hat{H}(B)$. Thus to complete the construction of the optimal ellipse it remains to consider all vertices of $C\hat{H}(B)$ other than Q in the first quadrant which lie in the region bounded by the spectral circle and Σ.[1] Let R be such a vertex and observe the following:

(i) By Lemma 3.2, the optimal capturing ellipse $\hat{E}_R$ through R has $\lambda_{\hat{E}_R} < 0$ and therefore excludes both Q and P.

(ii) Again by Lemma 3.2, the optimal capturing ellipse $\hat{E}_Q$ through Q must have a negative eccentricity and hence excludes P.

(iii) The optimal capturing ellipse through P is the line segment joining the points P and $(-\rho(B), 0)$ and therefore it excludes both Q and R.

(iv) The symmetric ellipse $E_{P,R}$, which passes through both P and R, excludes Q, as it passes through a point which lies in the interior of the unit circle.

(v) The symmetric ellipse $E_{Q,R}$, which passes through both Q and R, must have a real semi-axis greater than $\rho(B)$ and must contain P in its interior.

(vi) The symmetric ellipse $E_{P,Q}$, which passes through both P and Q, contains R.

From (ii), (v), and (vi) we see that

$$M_r(\hat{E}_Q) < \rho(B) = M_r(E_{P,Q}) < M_r(E_{Q,R}).$$

Consider the family $\mathcal{F}_Q$ of all capturing ellipses passing through the point Q and define the function a: $\mathcal{F}_Q \to [\alpha, 1)$ by

$$a = a(E_Q) = M_r(E_Q), \qquad E_Q \in \mathcal{F}_Q.$$

[1] For the possible locations of eigenvalues of nonnegative and stochastic matrices, see the paper by Dmitriev and Dynkin [1945].

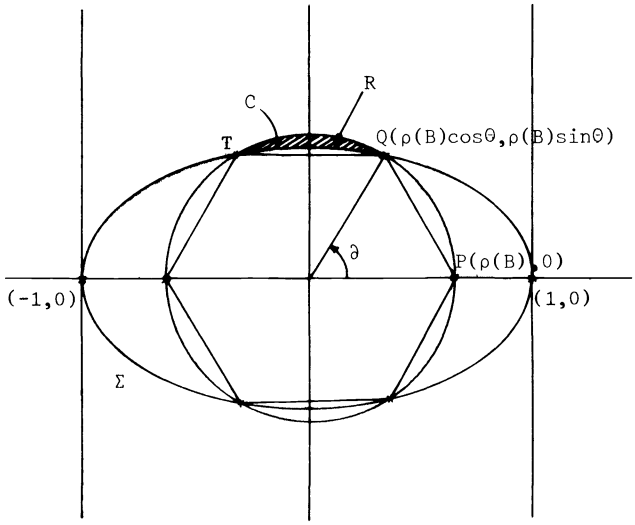


FIG. 2

Recall that for a given ellipse the imaginary axis can always be expressed in terms of its real axis and in terms of the coordinates of a point through which it passes. Hence we obtain, on identifying $\mu = \mu_{E_Q}$ with "ρ" in Young [1971, Chap. 6, eq. (4.27)], that

$$\mu_{E_{P,Q}} < \mu_{E_{Q,R}}.$$

Thus $E_{P,Q}$ is the optimal capturing ellipse for $C\hat{H}(B)$. Moreover, $E_{P,Q}$ is precisely the spectral circle $SC(B)$ and hence $\hat{\omega}_{opt} = 1$ by (2.10) because $\lambda_{E_{P,Q}} = 0$. This completes the proof. □

We now come to the consideration of the behaviour of the second-degree scheme (1.9) when B is 2-cyclic.

LEMMA 3.4. *Let B be an $n \times n$ nonnegative irreducible 2-cyclic matrix with $\rho(B) < 1$. Then the optimal iteration parameter $\hat{\omega}_{opt}$ for the iterative scheme* (1.9) *satisfies*

$$\omega_{opt} > 1.$$

Proof. By Romanovski's theorem (e.g., Varga [1962, Thm. 2.4]) with every $\lambda \in \sigma(B)$, $-\lambda \in \sigma(B)$. Also we know that as B is a real matrix, with every $\lambda \in \sigma(B)$, $\bar{\lambda} \in \sigma(B)$. Hence the spectrum of B is symmetric about the real and imaginary axes and so $C\hat{H}(B) = CH(B)$. Moreover, only two vertices of $C\hat{H}(B)$, $(\rho(B), 0)$ and $(-\rho(B), 0)$ lie on $SC(B)$.

It is a freshman calculus exercise to show that there exists an ellipse E, symmetric with respect to both axes, such that $M_r(E) = \rho(B)$ and $M_i(E) < \rho(B)$ and such that E is a capturing ellipse for $C\hat{H}(B)$. Thus, by (2.9),

$$\mu_E < \rho(B).$$

But then, by (2.8) and (2.10), $\hat{\omega}_{opt} \neq 1$. Notice that for E, $\omega = \omega(E) = 1 + \lambda_E \mu_E^2 > 1$. That $\hat{\omega}_{opt}$ must be greater than 1 is now an immediate consequence of Corollary 2.1. □

We remark that to find the optimal ω for the iteration scheme (1.9) when B is 2-cyclic we must follow the steps of the Young–Eidson [1970] algorithm as was shown by Avdelas, Galanis, and Hadjidimos [1983]. In the case when the spectrum of B is real, an explicit formula for $\hat{\omega}_{opt}$ has been obtained essentially by Golub and Varga [1961]

(see also Niethammer [1967]). It is the familiar expression

$$\hat{\omega}_{\text{opt}} = \frac{1}{1+\sqrt{1-\rho^2(B)}}. \tag{3.1}$$

To recoup: what we have accomplished so far in this section is to establish the behaviour of the iteration schemes (1.4) and (1.9) when the cyclicity of B is at least 2. We now come to the comparison of these schemes when $p = 1$, that is, when B is *primitive*. In this case we shall exhibit, by means of two examples, that whether scheme (1.4) or scheme (1.9) attains a superior rate of convergence depends on which part of the spectrum of B lies in the interior of SD (B).

Example 3.1. Let

$$B = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ .144 & .108 & .04 & .3 \end{bmatrix}. \tag{3.2}$$

Here $\sigma(B) = \{.8, -.5, \pm.6i\}$ and $\rho(B) = .8$. By following the analytical algorithm for determining the best extrapolation parameter for (1.4) of Hughes-Hallet [1981] or the geometrical algorithm obtained by Hadjidimos [1983], we can show that the optimal extrapolation parameter for B in (1.4) is $\omega_{\text{opt}} = 20/17 \approx 1.1765$, which yields

$$\rho(B_{\omega_{\text{opt}}}) = .7647.$$

To find the optimal iteration parameter $\hat{\omega}_{\text{opt}}$ for the second-order iteration scheme (1.9), we first note that in the present example $C\hat{H}(B)$ has only two vertices in the first quadrant $P_1(.8, 0)$ and $P_2(0, .6)$. In this case the application of the Young–Eidson [1970] algorithm results in an optimal capturing ellipse $\hat{E}$ with $M_r(\hat{E}) = .8$ and $M_i(\hat{E}) = .6$. Thus by (2.8)–(2.10), $\hat{\omega}_{\text{opt}} = 50(1 - .6\sqrt{2})/7 \doteq 1.0819$ and

$$\mu_{\text{opt}} = \mu_{\hat{E}} = 5(1 - .6\sqrt{2}) = .7574 < \rho(B_{\omega_{\text{opt}}}).$$

We therefore see that for B in (3.1) the second-degree scheme of (1.9) has a rate of convergence superior to that of the scheme in (1.4).

We comment that because of the spectral configuration of B in (3.2), we could have determined $\hat{\omega}_{\text{opt}}$ and μ_{opt} from Wrigley's [1963] formulas which coincide with equations (4.14) and (4.15) in Young [1971, Chap. 6].

Example 3.2. Let

$$B = \begin{bmatrix} .8 & .005 & 0 \\ .02 & .8 & .005 \\ 0 & .02 & .8 \end{bmatrix}. \tag{3.3}$$

Here $\sigma(B) = \{.8, .8 \pm .01\sqrt{2}\}$ and $\rho(B) = .8 + .01\sqrt{2}$. The fact that all eigenvalues of B are now real simplifies the problem of determining ω_{opt} for (1.4) and $\hat{\omega}_{\text{opt}}$ for (1.9). For (1.4) we see by (7) and (8) in Isaacson and Keller [1966, p. 76] that $\omega_{\text{opt}} = 5$, in which case

$$\rho(B_{\omega_{\text{opt}}}) = .05\sqrt{2} \approx .0707.$$

To determine the optimal iteration parameter for (1.9) we first notice that $C\hat{H}(B)$ has only one vertex in the first quadrant which is $(\rho(B), 0)$ and so the optimal capturing ellipse $\hat{E}$ for $C\hat{H}(B)$, and hence for $\sigma(B)$, has $M_r(\hat{E}) = \rho(B)$ and $M_i(\hat{E}) = 0$. From

(2.8)–(2.10) we obtain that $\hat{\omega}_{\text{opt}} = 2/(1 + \sqrt{1 - \rho^2(B)}) \approx 1.2653$ and

$$\mu_{\text{opt}} = \mu(\hat{E}) \approx .5151.$$

Thus $\rho(B_{\omega_{\text{opt}}}) < \mu_{\text{opt}}$ and we see that here, in contrast to Example 3.1, the extrapolated scheme (1.4) has an optimal rate of convergence superior to the optimal rate of convergence attained by the second degree of (1.9). We remark that the spectral configuration of B in (3.2) permits the determination of $\hat{\omega}_{\text{opt}}$ for the scheme (1.9) directly from the formula in (3.1) due to Golub and Varga [1961].

The results of the section are now summarized and included in the following theorem.

THEOREM 3.1. *Let $B \geqq 0$ be an $n \times n$, $n \geqq 2$, irreducible nonnegative matrix of cyclicity p and consider the three iteration schemes* (1.1), (1.4), *and* (1.9).

(i) *If $p \geqq 3$, then $\omega_{\text{opt}} = \hat{\omega}_{\text{opt}} = 1$ and*

$$\rho(B) = \rho(B_{\omega_{\text{opt}}}) = \mu_{\text{opt}}.$$

That is, no acceleration of the "original" iteration scheme (1.1) *can be achieved by either the extrapolation iteration scheme* (1.4) *or the second-degree stationary scheme* (1.9).

(ii) *If $p = 2$, then $\omega_{\text{opt}} = 1$, $\hat{\omega}_{\text{opt}} > 1$ and*

$$\mu_{\text{opt}} < \rho(B_{\omega_{\text{opt}}}) = \rho(B).$$

That is, an acceleration of the "original" iteration scheme (1.1) *is always possible by the second-degree scheme* (1.9), *but not by the extrapolation scheme* (1.4).

(iii) *If $p = 1$, then the "original" iteration scheme* (1.1) *can always be accelerated by either schemes* (1.4) *and* (1.9). *However, in general, which of the acceleration schemes* (1.4) *or* (1.9) *has a superior rate of convergence depends on the spectral configuration of B. In the special case when $\sigma(B)$ is real and ν is the minimal eigenvalue of B the following hold:*

$$\begin{aligned} &\rho(B_{\omega_{\text{opt}}}) < \mu_{\text{opt}}, && 1 - \sqrt{1 - \rho^2(B)} < \nu, \\ &\rho(B_{\omega_{\text{opt}}}) = \mu_{\text{opt}}, && 1 - \sqrt{1 - \rho^2(B)} = \nu, \\ &\mu_{\text{opt}} < \rho(B_{\omega_{\text{opt}}}), && \nu < 1 - \sqrt{1 - \rho^2(B)}. \end{aligned} \tag{3.4}$$

Proof. The proof of (i) is a consequence of Lemmas 3.1 and 3.3 and the fact that, when $\omega = 1$, both iterative schemes (1.4) and (1.9) reduce to the iterative scheme (1.1). The proof of (ii) is a consequence of Lemmas 3.1 and 3.4. We come now to the proof of (iii). Assume then that B satisfies the condition in (iii). The fact that (1.1) can always be accelerated by the extrapolation scheme (1.4) follows from Theorem 3.2 in de Pillis and Neumann [1981]; the details can be found in Hadjidimos [1986]. To see that (1.1) can always be accelerated by the second-degree scheme (1.9) we first note that, similar to the case $p = 2$ (see the proof of Lemma 3.4), there exists a capturing ellipse E of $C\hat{H}(B)$. The same is true for $\sigma(B)$, such that $M_r(E) = \rho(B)$ and $M_i(E) < \rho(B)$ showing, by (2.8)–(2.10), that $\mu_E < \rho(B)$. Next, that (1.4) does not consistently attain a rate of convergence superior to that of (1.9) and vice versa is an outcome of Examples 3.1 and 3.2.

Suppose now that $\sigma(B)$ is real. Then $C\hat{H}(B)$ is a line segment on the real axes, and hence it is its own optimal capturing ellipse $\hat{E}$, whence $M_r(\hat{E}) = \rho(B)$ and $M_i(\hat{E}) = 0$ so that by (2.9),

$$\mu_{\hat{E}} = \frac{\rho(B)}{1 + \sqrt{1 - \rho^2(B)}}. \tag{3.5}$$

Next, with ν being the smallest eigenvalue of B, it follows according to Isaacson and Keller [1966, pp. 75–76] that the optimal convergence rate of (1.4) occurs when

$$\omega_{\text{opt}} = \omega = \frac{2}{2 - (\rho(B) + \nu)}$$

and is given by

$$\rho(B_{\omega_{\text{opt}}}) = \frac{\rho(B) - \nu}{2 - (\rho(B) + \nu)}. \tag{3.6}$$

The various claims in (3.4) now follow from the comparison of (3.5) and (3.6). □

We conclude the paper by considering the acceleration of the iteration scheme (1.1) when B is an $n \times n$ nonnegative matrix with $\rho(B) < 1$, *but is reducible*. From (2.2) we know that

$$\sigma(B) = \bigcup_{j=1}^{k} \sigma(\tilde{B}_{jj}).$$

In the spirit of the terminology introduced by Rothblum [1975] we shall refer to a diagonal block in (2.2) as *basic* if

$$\rho(\tilde{B}_{jj}) = \rho(B).$$

It is the highest order of cyclicity among the basic blocks which largely governs whether the scheme (1.1) can be accelerated. Indeed let us now examine all possible cases.

Case 1. B has a basic block of cyclicity $p \geqq 3$. Then by following reasonings similar to those used in the proofs of Lemmas 3.1 and 3.3, we conclude that the convergence of (1.1) cannot be accelerated by either of the iteration schemes (1.4) or (1.9).

Case 2. B has a block of highest cyclicity $p = 2$. Then by following arguments similar to those used in the proof of Lemma 3.1, the convergence of (1.1) cannot be accelerated by the extrapolation scheme (1.4). As for the second-degree scheme of (1.9) it is clear that $C\hat{H}(B)$ can be captured by an ellipse E whose semi-axes satisfy

$$M_r(E) = \rho(B) > M_i(E),$$

which implies, by (2.8)–(2.10), that the scheme in (1.1) can be accelerated by (1.9). To find the optimal iteration parameter $\hat{\omega}_{\text{opt}}$ we must now apply the Young–Eidson [1970] algorithm.

Case 3. All basic blocks in B are primitive. By reasoning similar to that used in the proof of Theorem 3.1 (iii), the iteration scheme (1.1) can always be accelerated by both the extrapolation scheme (1.4) and the second-degree scheme (1.9). As in the case when B itself is irreducible and primitive, it is not generally possible to determine a priori which scheme will provide the superior optimal rate of convergence. To find the best extrapolation parameter ω_{opt} we must follow the algorithms developed by Hughes-Hallet [1981] or by Hadjidimos [1983]. To find the optimal iteration parameter $\hat{\omega}_{\text{opt}}$ for the second-degree scheme (1.9) we must once again follow the algorithm due to Young and Eidson [1970].

REFERENCES

G. AVDELAS, S. GALANIS, AND A. HADJIDIMOS [1983], *On the optimization of a class of second order iterative schemes*, BIT, 23, pp. 50–64.

G. AVDELAS AND A. HADJIDIMOS [1983], *Optimum second order stationary extrapolated iterative schemes*, Math. Comput. Simulation, 25, pp. 189–198.

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

J. DE PILLIS [1980], *How to embrace your spectrum for faster iterative results*, Linear Algebra Appl., 34, pp. 125–143.

J. DE PILLIS AND M. NEUMANN [1981], *Iterative methods with k-part splittings*, IMA J. Numer. Anal., 1, pp. 65–79.

N. DMITRIEV AND E. DYNKIN [1945], *On the characteristic numbers of a stochastic matrix*, Dokl. Akad. Nauk SSSR, 3, pp. 159–163.

G. H. GOLUB AND R. S. VARGA [1961], *Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second-order Richardson iterative methods*, Numer. Math., 3, pp. 147–168.

A. HADJIDIMOS [1983], *The optimal solution of the extrapolation problem of a first order scheme*, Internat. J. Comput. Math., 13, pp. 153–168.

——— [1986], *On the extrapolation technique for the solution of linear systems*, Calcolo, 23, pp. 35–43.

A. J. HUGHES-HALLETT [1981], *Some extensions and comparisons in the theory of Gauss–Seidel iterative techniques for solving large equation systems*, in Proc. Econometric Society European Meeting 1979, E. G. Charatsis, ed., North–Holland, Amsterdam, pp. 279–318.

E. ISAACSON AND M. B. KELLER [1966], *Analysis of Numerical Methods*, John Wiley, New York.

U. KULISCH [1968], *Über regulare Zerlegungen von Matrizen und einige Anwendungen*, Numer. Math., 11, pp. 444–449.

W. NIETHAMMER [1967], *Iterationsverfahren und allgemeine Euler-Verfahren*, Math. Zeit., 102, pp. 288–317.

U. ROTHBLUM [1975], *Algebraic eigenspaces of non-negative matrices*, Linear Algebra Appl., 12, pp. 281–292.

R. S. VARGA [1962], *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, N.J.

E. E. WRIGLEY [1962], *On accelerating the Jacobi method for solving simultaneous equations by Chebyshev extrapolation when the eigenvalues of the iteration matrix are complex*, Comput. J., 6, pp. 169–176.

D. M. YOUNG [1971], *Iterative Solutions of Large Linear Systems*, Academic Press, New York.

D. M. YOUNG AND M. E. EIDSON [1970], *On the determination of optimum relaxation factor for the SOR method when the eigenvalues of the Jacobi Matrix are complex*, Report CNA-1, Center for Numerical Analysis, Univ. of Texas, Austin, TX.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

005

ON THE RANGE OF THE HADAMARD PRODUCT OF A POSITIVE DEFINITE MATRIX AND ITS INVERSE*

MIROSLAV FIEDLER†‡ AND THOMAS L. MARKHAM‡

Abstract. Suppose A is an $n \times n$ real positive definite matrix. This paper characterizes for $n = 3$ the range of the operator $A \circ A^{-1}$, where $\circ$ denotes the Hadamard product. It is shown that the necessary conditions exhibited by the first author in 1964 [M. Fiedler, *Czechoslovak Math. J.*, 14 (1964), pp. 39–51] are also sufficient in this case.

Key words. real positive definite matrix, Hadamard product, nonnegative matrix

AMS(MOS) subject classifications. 15A23, 15A45

In [2], it is shown that if A is a positive definite matrix, then the Hadamard product $P = A \circ (A^{-1})^T$, has the property that the matrix $P - I$ is positive semidefinite and $Pe = e$, where e is the column vector of all ones. In [1], it is even shown that the multiplicity of 1 as the eigenvalue of P is equal to the maximum number of diagonal blocks in all possible block diagonal forms to which the matrix A can be brought by simultaneous permutations of rows and columns.

Then, in [3], the following necessary and sufficient condition for the diagonal entries p_{ii} of such a matrix P is found:

$$2 \max_i (p_{ii}^{1/2} - 1) \leqq \sum_{i=1}^{n} (p_{ii}^{1/2} - 1).$$

In the present paper, we shall show that in the 3-by-3 case, the just-stated necessary conditions for a real matrix P to have the form $A \circ A^{-1}$, A real symmetric and positive definite, are also sufficient. The explicit form of A is also given.

THEOREM. *Let $P = (p_{ij})$ be a real symmetric 3-by-3 matrix. Then the following are equivalent*:

(i) *$P = A \circ A^{-1}$ for some real symmetric positive definite matrix A.*

(ii) *$P - I$ is positive semidefinite, $Pe = e$ (e is again $(1, 1, \cdots 1)^T$) and*

$$2 \max_i (p_{ii}^{1/2} - 1) \leqq \sum_{i=1}^{3} (p_{ii}^{1/2} - 1). \tag{1}$$

(iii)

$$P = \begin{bmatrix} 1+p_2+p_3 & -p_3 & -p_2 \\ -p_3 & 1+p_1+p_3 & -p_1 \\ -p_2 & -p_1 & 1+p_1+p_2 \end{bmatrix} \tag{2}$$

for some real numbers p_1, p_2, p_3 satisfying

$$p_2 + p_3 \geqq 0, \quad p_1 + p_3 \geqq 0, \quad p_1 + p_2 \geqq 0, \tag{3}$$

$$p_1p_2 + p_1p_3 + p_2p_3 \geqq 0, \tag{4}$$

$$(p_1p_2 + p_1p_3 + p_2p_3)^2 + 4p_1p_2p_3 \geqq 0. \tag{5}$$

* Received by the editors June 8, 1987; accepted for publication (in revised form) September 30, 1987.
† Mathematics Institute, Czechoslovak Academy of Sciences, Prague, Czechoslovakia.
‡ Department of Mathematics, University of South Carolina, Columbia, South Carolina 29208.

(iv) $P = A \circ A^{-1}$, *where* A *is an elementwise nonnegative and real symmetric positive definite matrix.*

In terms of condition (iii), *the matrix* A *can be chosen as that nonnegative matrix*

$$\begin{bmatrix} 1 & a_{12} & a_{13} \\ a_{12} & 1 & a_{23} \\ a_{13} & a_{23} & 1 \end{bmatrix} \tag{6}$$

for which

$$a_{12}^2 = \frac{2p_3 + Q + W}{2 + 2S + Q + W},$$

$$a_{13}^2 = \frac{2p_2 + Q + W}{2 + 2S + Q + W}, \tag{7}$$

$$a_{23}^2 = \frac{2p_1 + Q + W}{2 + 2S + Q + W},$$

where

$$S = p_1 + p_2 + p_3, \qquad Q = p_1p_2 + p_1p_3 + p_2p_3 \tag{8}$$

and W *is the nonnegative square root of the left-hand side in* (5).

Proof. The implication (i) → (ii) was proved in [3].

We have that the implication (ii) → (iii). If (ii) is fulfilled, P clearly has the form (2) with real numbers p_i satisfying (3) and (4). It remains to prove (5). If p_i are all nonnegative, (5) is fulfilled. Thus, let $p_3 < 0$, say, and denote

$$|p_3| = m, \quad \omega_1 = p_1 + p_3, \quad \omega_2 = p_2 + p_3. \tag{9}$$

Then

$$\omega_1 \geqq 0, \qquad \omega_2 \geqq 0, \tag{10}$$

and by (4),

$$\omega_1\omega_2 \geqq m^2. \tag{11}$$

Condition (1) then reads

$$(1 + 2m + \omega_1 + \omega_2)^{1/2} + 1 \leqq (1 + \omega_1)^{1/2} + (1 + \omega_2)^{1/2}; \tag{1'}$$

condition (5), which we have to prove, becomes

$$(\omega_1\omega_2 - m^2)^2 - 4m(m + \omega_1)(m + \omega_2) \geqq 0. \tag{5'}$$

Now, (1′) implies

$$1 + 2m + \omega_1 + \omega_2 + 1 + 2(1 + 2m + \omega_1 + \omega_2)^{1/2} \leqq 1 + \omega_1 + 1 + \omega_2 + 2(1 + \omega_1)^{1/2}(1 + \omega_2)^{1/2},$$

which is equivalent to

$$m + (1 + 2m + \omega_1 + \omega_2)^{1/2} \leqq (1 + \omega_1)^{1/2}(1 + \omega_2)^{1/2}.$$

Again, this implies

$$m^2 + 1 + 2m + \omega_1 + \omega_2 + 2m(1 + 2m + \omega_1 + \omega_2)^{1/2} \leqq 1 + \omega_1 + \omega_2 + \omega_1\omega_2,$$

which can be rewritten as

$$2m(1 + 2m + \omega_1 + \omega_2)^{1/2} \leqq \omega_1\omega_2 - m^2 - 2m.$$

Both sides being nonnegative, it follows that

$$4m^2(1+2m+\omega_1+\omega_2) \leqq 4m^2+4m(m^2-\omega_1\omega_2)+(\omega_1\omega_2-m^2)^2.$$

The left-hand side being

$$4m^2+4m(m^2-\omega_1\omega_2)+4m(m+\omega_1)(m+\omega_2),$$

we obtain the inequality (5′).

We have that the implication (iii) → (iv). For P given in (2) and p_i satisfying (3)–(5), we have to show the following:

(a) The right-hand sides in (7) are nonnegative (the denominators are clearly positive);

(b) The matrix A defined by (6) is positive definite;

(c) $P = A \circ A^{-1}$.

To prove (a), suppose that, say,

$$2p_3+Q+W<0.$$

Then,

$$2p_3<-(Q+W) \leqq 0$$

and by (3), $p_1 \geqq |p_3|$, $p_2 \geqq |p_3|$, so that by (5),

$$(p_1p_2+p_1p_3+p_2p_3)^2 \geqq 4p_1p_2|p_3| > 2p_1p_2(Q+W) \geqq 2p_1p_2Q.$$

Since $Q \geqq 0$, we have

$$Q > 2p_1p_2,$$

which implies

$$p_3(p_1+p_2) > p_1p_2,$$

a contradiction.

To prove (b), we shall show that all principal minors of A are positive. The diagonal entries are positive. Now,

$$1-a_{12}^2=\frac{2+2(p_1+p_2)}{2+2S+Q+W} \tag{12}$$

implies this for minors of degree two by (3). The determinant Δ of A is

$$\Delta = 1-a_{12}^2-a_{13}^2-a_{23}^2+2R,$$

where

$$R = a_{12}a_{13}a_{23}. \tag{13}$$

By (7),

$$R^2=\frac{8p_1p_2p_3+4Q(Q+W)+2S(Q+W)^2+(Q+W)^3}{(2+2S+Q+W)^3}.$$

Since $4p_1p_2p_3 = W^2 - Q^2$,

$$\begin{aligned} R^2 &= \frac{Q+W}{(2+2S+Q+W)^3}(2(W-Q)+4Q+2S(Q+W)+(Q+W)^2) \\ &= \frac{(Q+W)^2}{(2+2S+Q+W)^3}(2+2S+Q+W) \end{aligned}$$

so that

$$R=(Q+W)/(2+2S+Q+W). \tag{14}$$

Again by (7),

$$a_{12}^2+a_{13}^2+a_{23}^2=(2S+3Q+3W)/(2+2S+Q+W),$$

which easily implies

$$\Delta=(2+2S+Q+W-(2S+3Q+3W)+2(Q+W))/(2+2S+Q+W),$$

i.e.,

$$\Delta=2/(2+2S+Q+W). \tag{15}$$

Thus $\Delta > 0$.

To prove (c), observe that

$$A^{-1}=\frac{1}{\Delta}\begin{bmatrix} 1-a_{23}^2 & a_{13}a_{23}-a_{12} & a_{12}a_{23}-a_{13} \\ a_{13}a_{23}-a_{12} & 1-a_{13}^2 & a_{12}a_{13}-a_{23} \\ a_{12}a_{23}-a_{13} & a_{12}a_{13}-a_{23} & 1-a_{12}^2 \end{bmatrix}$$

so that by (13),

$$A\circ A^{-1}=\frac{1}{\Delta}\begin{bmatrix} 1-a_{23}^2 & R-a_{12}^2 & R-a_{13}^2 \\ R-a_{12}^2 & 1-a_{13}^2 & R-a_{23}^2 \\ R-a_{13}^2 & R-a_{23}^2 & 1-a_{12}^2 \end{bmatrix}.$$

It then follows easily from (7), (12), (14), and (15) that $P = A\circ A^{-1}$. Since the implication (iv) $\rightarrow$ (i) is trivial, the proof is complete. □

For $n \geqq 4$, conditions (i) and (ii) of our theorem are not equivalent. To understand the reason, we return to a concept used in [4].

DEFINITION [4]. Let A be an $n \times n$ matrix over a field F. The off-diagonal rank of A is the smallest integer w, with the property that there exists a $w \times w$ nonsingular submatrix of A which does not contain any diagonal entry of A, and every submatrix of A of order $w + 1$ is either singular or must contain a diagonal entry of A.

In [3, Thm. 3.3, p. 46], the conditions for equality to hold in (ii)–(1) are given. In the case of equality, it is straightforward to see that P must have off-diagonal rank less than or equal to 1.

Now consider

$$P=I+\begin{bmatrix} 3 & -\frac{8}{9} & -\frac{8}{9} & -\frac{11}{9} \\ -\frac{8}{9} & \frac{7}{9} & -\frac{1}{9} & \frac{2}{9} \\ -\frac{8}{9} & -\frac{1}{9} & \frac{7}{9} & \frac{2}{9} \\ -\frac{11}{9} & \frac{2}{9} & \frac{2}{9} & \frac{7}{9} \end{bmatrix}.$$

P has the property that $Pe = e$, $P - I$ is positive semidefinite, and equality holds in condition (ii)–(1) of the theorem. However, P has off-diagonal rank of 2, so $P \neq A\circ A^{-1}$ for any positive definite matrix A.

For $n > 4$, we obtain the same conclusion by considering the matrix $P \oplus I_{n-4}$, where I_{n-4} denotes the identity matrix of order $n - 4$.

Observation. The range of the mapping $A \rightarrow A\circ A^{-1}$ is not convex.

The matrices

$$M_1 = \begin{bmatrix} \frac{289}{81} & \frac{8}{81} & -\frac{8}{3} \\ \frac{8}{81} & \frac{100}{81} & -\frac{1}{3} \\ -\frac{8}{3} & -\frac{1}{3} & 4 \end{bmatrix},$$

$$M_2 = \begin{bmatrix} 25 & 16 & -40 \\ 16 & 25 & -40 \\ -40 & -40 & 81 \end{bmatrix}$$

belong to the range.
Indeed, $M_1 = A_1 \circ A_1^{-1}$, where

$$A_1 = \begin{bmatrix} \frac{17}{9} & \frac{2\sqrt{2}}{9} & \frac{2\sqrt{6}}{3} \\ \frac{2\sqrt{2}}{9} & \frac{10}{9} & \frac{\sqrt{3}}{3} \\ \frac{2\sqrt{6}}{3} & \frac{\sqrt{3}}{3} & 2 \end{bmatrix}$$

and

$$M_2 = A_2 \circ A_2^{-1},$$

where

$$A_2 = \begin{bmatrix} 5 & 4 & 2\sqrt{10} \\ 4 & 5 & 2\sqrt{10} \\ 2\sqrt{10} & 2\sqrt{10} & 9 \end{bmatrix}.$$

However, the matrix $(1 - \varepsilon)M_1 + \varepsilon M_2$, for sufficiently small positive ε, does not belong to the range since its diagonal entries

$$\tfrac{289}{81} + \tfrac{1736}{81}\varepsilon, \quad \tfrac{100}{81} + \tfrac{1925}{81}\varepsilon, \quad 4 + 77\varepsilon$$

do not satisfy condition (ii)–(1). This condition then reads, as $O(\varepsilon^2)$,

$$3 + \tfrac{77}{4}\varepsilon \leqq \left[\tfrac{17}{9} + \tfrac{868}{153}\varepsilon\right] + \left[\tfrac{10}{9} + \tfrac{1925}{180}\varepsilon\right],$$

which is false for ε small.

We note that this problem of characterizing the range of $A \circ A^{-1}$ has been mentioned in [5].

REFERENCES

[1] M. FIEDLER, *On some properties of hermitian matrices*, (Czechoslovak) Mat.-fyz. Cas. SAV, 7 (1957), pp. 168–176.

[2] ———, *Uber eine Ungleichung fur positiv definite Matrizen*, Math. Nachr., 23 (1961), pp. 197–199.

[3] ———, *Relations between the diagonal elements of two mutually inverse positive definite matrices*, Czechoslovak Math. J., 14 (1964), pp. 39–51.

[4] M. FIEDLER AND T. L. MARKHAM, *Rank-preserving diagonal completions of a matrix*, Linear Algebra Appl., 85 (1987), pp. 49–56.

[5] C. R. JOHNSON AND H. SHAPIRO, *Mathematical aspects of the relative gain array* ($A \circ A^{-T}$), SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 627–644.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

006

THE MATRIX EQUATION $A\bar{X} - XB = C$ AND ITS SPECIAL CASES*

JEAN H. BEVIS†, FRANK J. HALL†, AND ROBERT E. HARTWIG‡

Abstract. The consistency and solutions of the matrix equations $A\bar{X} - XB = C$, $A\bar{X} \pm XA^T = C$, and $A\bar{X} \pm XA^* = C$ are characterized. As a consequence it is shown that A^T (respectively, A^*) may be obtained from A by a consimilarity transformation using a Hermitian (respectively, symmetric) matrix.

Key words. matrix equation, consimilarity, conjugation

AMS(MOS) subject classifications. 15A24

1. Introduction. This paper extends the results of [2] concerning the matrix equation $A\bar{X} - XB = C$, where A and B are complex m-by-m and n-by-n matrices, respectively, and $\bar{X}$ denotes the matrix obtained by taking the complex conjugate of each element of X. The equation is said to be consistent for given matrices A, B, and C if there is a matrix X for which the equality holds. In § 2 we give closed-form consistency conditions and particular solutions for this equation. The general solution to the corresponding homogeneous equation $A\bar{X} - XB = O$ is given in [2], where it is noted that any solution to the equation $A\bar{X} - XB = C$ can be written as a particular solution plus a complementary solution to the homogeneous equation. Thus we also obtain the general solution to $A\bar{X} - XB = C$.

Square matrices A and B are said to be consimilar if there is a nonsingular matrix P such that $A = PB\bar{P}^{-1}$. Consimilarity first appeared as a change of basis for the matrix representation of semilinear transformations [8], [9]. More recent interest in consimilarity arises in the work of Hong and Horn [4]–[6], and in the study of matrix products of the form $UAU^T = UA\bar{U}^{-1}$, where U is complex unitary and U^T denotes the transpose of U [7, § 4.6].

A consimilarity version of Roth's theorem is given in [2], namely

$$\begin{bmatrix} A & C \\ O & B \end{bmatrix}$$

is consimilar to

$$\begin{bmatrix} A & O \\ O & B \end{bmatrix}$$

if and only if $A\bar{X} - XB = C$ has a solution; furthermore, if X is a solution to this equation, then the consimilarity may be carried out via the matrix

$$\begin{bmatrix} I & X \\ O & I \end{bmatrix}.$$

Thus the results of § 2 may be of interest in simplifying matrix representations of semilinear transformations arising from quantum mechanics.

In § 3 we show that the results of § 2 are sufficient to characterize solutions to the matrix equations $A\bar{X} \pm XA^T = C$ and $A\bar{X} \pm XA^* = C$ where A^* denotes the conjugate transpose of A. It is known [5] that A is consimilar to A^T and to A^*. As an application

* Received by the editors June 15, 1987; accepted for publication (in revised form) September 16, 1987.
† Department of Mathematics and Computer Science, Georgia State University, Atlanta, Georgia 30303.
‡ Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695.

of the results of § 3 we show that A is consimilar to A^T via a Hermitian matrix and A is consimilar to A^* via a symmetric matrix.

2. The equation $A\bar{X} - XB = C$. In order to simplify the equation $A\bar{X} - XB = C$ we use a canonical form introduced by Hong and Horn [5] who showed that any square complex matrix A is consimilar to a direct sum of blocks of the form $J_k(\lambda)$ with $\lambda \geqq 0$, and

$$\begin{bmatrix} O & I_k \\ J_k(\lambda) & O \end{bmatrix}$$

with $\lambda < 0$ or $\operatorname{Im}(\lambda) > 0$, where $\operatorname{Im}(\lambda)$ is the imaginary part of λ, I_k is the k-by-k identity, and $J_k(\lambda)$ is the k-by-k Jordan block

$$\begin{bmatrix} \lambda & 1 & & O \\ & \lambda & 1 & \\ & & \ddots & \\ O & & & 1 \\ & & & \lambda \end{bmatrix}.$$

Such a direct sum is called a concanonical form for A and is also discussed in [2]. The set of λ appearing in the $J_k(\lambda)$ of these blocks is called the conspectrum of A and is denoted by $c\sigma(A)$. Now $c\sigma(A)$ may be determined from $\sigma(A\bar{A})$, the spectrum of $A\bar{A}$. Moreover, according to [2, Thm. 3] $A\bar{X} - XB = C$ has a unique solution for all C if and only if $c\sigma(A) \cap c\sigma(B)$ is empty, which is equivalent to $\sigma(A\bar{A}) \cap \sigma(B\bar{B})$ being empty.

Let P and Q be nonsingular matrices such that $K = PA\bar{P}^{-1}$ and $L = QB\bar{Q}^{-1}$ are the concanonical forms of A and B, respectively. Let $Y = PXQ^{-1} = [Y_{st}]$ and $PC\bar{Q}^{-1} = [C_{st}]$ have rows partitioned according to the blocks of K and columns partitioned according to the blocks of L. Then $A\bar{X} - XB = C$ is equivalent to $K\bar{Y} - YL = PC\bar{Q}^{-1}$ which is equivalent to $K_s\bar{Y}_{st} - Y_{st}L_t = C_{st}$ for all s, t, where K_s and L_t denote diagonal blocks of K and L, respectively. Indeed there are four possible forms for the last equation depending on the form of the blocks K_s and L_t. These are as follows.

Case I. $J_m(\lambda)\bar{Y} - YJ_n(\mu) = C$ with $\lambda, \mu \geqq 0$.

Case II.

$$\begin{bmatrix} O & I_m \\ J_m(\lambda) & O \end{bmatrix}\bar{Y} - YJ_n(\mu) = C \quad \text{with } \lambda < 0 \text{ or } \operatorname{Im}(\lambda) \neq 0, \text{ and } \mu \geqq 0.$$

Case III.

$$J_m(\lambda)\bar{Y} - Y\begin{bmatrix} O & I_n \\ J_n(\mu) & O \end{bmatrix} = C \quad \text{with } \lambda \geqq 0, \text{and } \mu < 0 \text{ or } \operatorname{Im}(\mu) \neq 0.$$

Case IV.

$$\begin{bmatrix} O & I_m \\ J_m(\lambda) & O \end{bmatrix}\bar{Y} - Y\begin{bmatrix} O & I_n \\ J_n(\mu) & O \end{bmatrix} = C \quad \text{with } \lambda < 0 \text{ or } \operatorname{Im}(\lambda) \neq 0, \text{and } \mu < 0 \text{ or } \operatorname{Im}(\mu) \neq 0.$$

These four cases and their possible subcases will be treated in Lemmas 2–5 herein. Our results will be based on results of [3] and [10] which are summarized in Lemma 1.

In order to portray the relationships between the solutions to these cases we introduce the following matrix functions. For an m-by-n matrix C and conformal matrices

U, V let

$$\Phi_L(U,C,V)=\sum_{k=0}^{m-1} U^kCV^{-k-1}, \qquad \Phi_R(U,C,V)=\sum_{k=0}^{n-1} U^{-k-1}CV^k,$$

$$\Omega_L(C)=\sum_{k=0}^{\nu-1} J_m^kC(J_n^T)^k, \qquad \Omega_R(C)=\sum_{k=0}^{\nu-1} (J_m^T)^kCJ_n^k,$$

$$\Gamma_L(C)=\sum_{k=0}^{\nu-1} J_m^k\gamma^k(C)(J_n^T)^k, \qquad \Gamma_R(C)=\sum_{k=0}^{\nu-1} (J_m^T)^k\gamma^k(C)J_n^k$$

where $J_m = J_m(0)$, $\nu = \min(m, n)$, and $\gamma(C) = [\bar{c}_{ij}]$ when $C = [c_{ij}]$. Thus $\gamma^k(C) = C$ for k even and $\gamma^k(C) = \bar{C}$ for k odd. We use several elementary properties of these functions, such as $\Phi_L(-U, C, V) = -\Phi_L(U, C, -V)$ and $\Phi_L(U, CS, V)S^{-1} = \Phi_L(U, C, SVS^{-1})$, in the computations below, where we also use e_k to denote the kth column of the appropriately sized identity matrix.

LEMMA 1. *Consider the equation* $J_m(\lambda)Y - YJ_n(\mu) = C$.

(i) *If* $\lambda \neq \mu$ *then there is a unique solution*

$$Y=-\Phi_L[J_m,C,J_n(\mu-\lambda)]=\Phi_R[J_m(\lambda-\mu),C,J_n]. \tag{2.1}$$

(ii) *If* $\lambda = \mu$, *then the equation has a solution if and only if* C *satisfies the consistency condition*

$$\Omega_L(C)e_1=O \quad \text{for } m\leqq n \quad \text{or} \quad e_m^T\Omega_R(C)=O \quad \text{for } m\geqq n. \tag{2.2}$$

In that case a particular solution is given by

$$Y_p=-\Omega_L(C)J_n^T \quad \text{for } m\leqq n \quad \text{or} \quad Y_p=J_m^T\Omega_R(C) \quad \text{for } m\geqq n, \tag{2.3}$$

and the general solution is $Y = Y_p + Y_c$, *where* Y_c *is an arbitrary upper Toeplitz matrix.*

An m-by-n matrix $Y = (y_{ij})$ is said to be Toeplitz (respectively, conjugate Toeplitz) if $y_{i+1,j+1} = y_{i,j}$ (respectively, $y_{i+1,j+1} = \bar{y}_{i,j}$) for all $i < m$ and $j < n$. We say that a matrix is upper if it has the form $[O\ T]$, T, or $\begin{bmatrix} T \\ O \end{bmatrix}$, where T is a square upper triangular matrix. Upper Toeplitz and upper conjugate Toeplitz matrices arise in the solution of the equations $J_mY - YJ_n = O$ and $J_m\bar{Y} - YJ_n = O$, respectively. Conjugate Toeplitz matrices were introduced in [1] as solutions to similar equations involving companion matrices. Below we denote the real and imaginary parts of a matrix C by Re (C) and Im (C), respectively. We are now ready for our first case, which has three subcases.

Case I. $\lambda \geqq 0$, $\mu \geqq 0$.

LEMMA 2. *Consider the equation* $J_m(\lambda)\bar{Y} - YJ_n(\mu) = C$, *where* $\lambda, \mu \geqq 0$, *and* C *is an* m*-by-*n *complex matrix.*

(i) *If* $\lambda \neq \mu$ *then the equation has a unique solution given by*

$$\begin{aligned} Y&=-\Phi_L[J_m,\operatorname{Re}(C),J_n(\mu-\lambda)]-i\Phi_L[-J_m,\operatorname{Im}(C),J_n(\lambda+\mu)]\\ &=\Phi_R[J_m(\lambda-\mu),\operatorname{Re}(C),J_n]-i\Phi_R[J_m(\lambda+\mu),\operatorname{Im}(C),-J_n]. \end{aligned} \tag{2.4}$$

(ii) *If* $\lambda = \mu > 0$ *then a solution exists if and only if*

$$\Omega_L(\operatorname{Re}(C))e_1=O \quad \text{for } m\leqq n \quad \text{or} \quad e_m^T\Omega_R(\operatorname{Re}(C))=O \quad \text{for } m\geqq n. \tag{2.5}$$

In this case a particular solution is given by

$$\begin{aligned} &Y_p=-\Omega_L[\operatorname{Re}(C)]J_n^T-i\Phi_L[-J_m,\operatorname{Im}(C),J_n(2\lambda)] \quad \text{for } m\leqq n, \quad \text{or}\\ &Y_p=J_m^T\Omega_R[\operatorname{Re}(C)]-i\Phi_R[J_m(2\lambda),\operatorname{Im}(C),-J_n] \quad \text{for } m\geqq n \end{aligned} \tag{2.6}$$

and the general solution is given by $Y = Y_p + Y_c$, *where* Y_c *is an arbitrary real upper Toeplitz matrix.*

(iii) *If* $\lambda = \mu = 0$ *then a solution exists if and only if*

$$\Gamma_L(C)e_1 = O \quad \text{for } m \leqq n \quad \text{or} \quad e_m^T \Gamma_R(\bar{C}) = O \quad \text{for } m \geqq n. \tag{2.7}$$

In this case a particular solution is given by

$$Y_p = -\Gamma_L(C)J_n^T \quad \text{for } m \leqq n \quad \text{or} \quad Y_p = J_m^T \Gamma_R(\bar{C}) \quad \text{for } m \geqq n \tag{2.8}$$

and the general solution is given by $Y = Y_p + Y_c$ *where* Y_c *is an arbitrary upper conjugate Toeplitz matrix.*

Proof. By writing $Y = U + iV$ and $C = D + iE$ in terms of their real and imaginary parts we see that the equation is equivalent to

$$J_m(\lambda)U - UJ_n(\mu) = D, \tag{2.9}$$

$$J_m(\lambda)V + VJ_n(\mu) = -E. \tag{2.10}$$

Let H be the n-by-n diagonal matrix $H = H^{-1} = \operatorname{diag}(1, -1, 1, -1, \cdots)$. Since $HJ_n(\mu)H = -J_n(-\mu)$, (2.10) is equivalent to

$$J_m(\lambda)VH - VHJ_n(-\mu) = -EH. \tag{2.11}$$

If $\lambda \neq \mu$ are nonnegative, then $\lambda \neq -\mu$ so that (2.9) and (2.11) have unique solutions. The real part of Y in (2.4) follows directly from (2.1) which also implies that

$$\begin{aligned} V &= -\Phi_L[J_m, -EH, J_n(-\mu - \lambda)]H = \Phi_R[J_m(\lambda + \mu), -EH, J_n]H \\ &= \Phi_L[J_m, E, HJ_n(-\lambda - \mu)H] = -\Phi_R[J_m(\lambda + \mu), E, HJ_nH] \\ &= -\Phi_L[-J_m, E, J_n(\lambda + \mu)] = -\Phi_R[J_m(\lambda + \mu), E, -J_n]. \end{aligned}$$

This gives the imaginary parts of Y in (2.4), and for case (ii), where $\lambda \neq -\mu$, it also gives the imaginary parts of Y in (2.6).

If $\lambda = \mu > 0$, then $\lambda \neq -\mu$, so there is a unique solution for the imaginary part of Y, as we just noted. However, (2.9) may or may not be consistent. Thus (2.5) and the real parts of (2.6) are obtained by applying (2.2) and (2.3) to (2.9).

If $\lambda = \mu = 0$, then the real parts of (2.7) and (2.8) are obtained by applying (2.2) and (2.3) to (2.9) as above. Also $\lambda = -\mu$ so that (2.11) is consistent if and only if $O = \Omega_L(-EH)e_1$ for $m \leqq n$ or $O = e_m^T \Omega_R(-EH)$ for $m \geqq n$. Note that for real matrices such as E, $\Omega_L(iEH)H = \Gamma_L(iE)$ and $\Omega_R(iEH)H = \Gamma_R(iE)$. Thus the consistency condition for (2.11) is equivalent to

$$O = [\Omega_L(iEH)H][He_1] = \Gamma_L(iE)e_1 \quad \text{for } m \leqq n, \quad \text{or}$$

$$O = e_m^T \Omega_R(-iEH)H = e_m^T \Gamma_R(-iE) \quad \text{for } m \geqq n$$

which gives the imaginary parts of (2.7). If (2.11) is consistent, then a particular solution for VH is given by

$$V_pH = -\Omega_L(-EH)J_n^T \quad \text{for } m \leqq n \quad \text{or} \quad V_pH = J_m^T \Omega_R(-EH) \quad \text{for } m \geqq n,$$

or equivalently,

$$iV_p = [\Omega_L(iEH)H][HJ_n^T H] = -\Gamma_L(iE)J_n^T \quad \text{for } m \leqq n, \quad \text{or}$$

$$iV_p = J_m^T \Omega_R(-iEH)H = J_m^T \Gamma_R(\bar{iE}) \qquad \text{for } m \geqq n.$$

This establishes the imaginary parts of (2.8). The complementary solutions are given by

Theorem 2(2), 2(3) of [2].

Case II. $\lambda < 0$ or $\operatorname{Im}(\lambda) \neq 0$, and $\mu \geqq 0$.

LEMMA 3. *Consider the equation*

$$\begin{bmatrix} O & I_m \\ J_m(\lambda) & O \end{bmatrix} \bar{Y} - YJ_n(\mu) = C$$

where $\lambda < 0$ *or* $\operatorname{Im}(\lambda) \neq 0$, $\mu \geqq 0$, *and* C *is a* $2m$*-by-*n *matrix. Partition* Y *and* C *conformally into* m*-by-*n *submatrices as*

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}.$$

Then there is a unique solution of the form

$$Y = \begin{bmatrix} Y_1 \\ \bar{Y}_1 J_n(\mu) + \bar{C}_1 \end{bmatrix} \quad \text{with } Y_1 = \Phi_L[-J_m, \bar{C}_2 + C_1 J_n(\mu), \bar{\lambda} I_n - J_n^2(\mu)]. \tag{2.12}$$

Proof. The conspectrum of

$$\begin{bmatrix} O & I_m \\ J_m(\lambda) & O \end{bmatrix}$$

and $J_n(\mu)$ are disjoint, so by Theorem 3 of [2] the equation has a unique solution. When the equation is written in terms of the submatrices of Y and C we obtain

$$J_m(\lambda)\bar{Y}_1 - Y_2 J_n(\mu) = C_2 \tag{2.13}$$

and $\bar{Y}_2 - Y_1 J_n(\mu) = C_1$. Thus $Y_2 = \bar{Y}_1 J_n(\mu) + \bar{C}_1$, which gives the form of (2.12) and when combined with (2.13) gives $J_m(\bar{\lambda})Y_1 - Y_1 J_n^2(\mu) = W$, where $W = \bar{C}_2 + C_1 J_n(\mu)$. Since the spectra of $J_m(\bar{\lambda})$ and $J_n^2(\mu)$ are disjoint ($\bar{\lambda} \neq \mu^2$), the last equation uniquely determines Y_1. This equation may be solved by iteration. Indeed, since $J_m(\bar{\lambda}) = J_m + \bar{\lambda} I_m$, we have $J_m Y_1 + Y_1[\bar{\lambda} I_n - J_n^2(\mu)] = W$, and hence

$$Y_1 = [W - J_m Y_1][\bar{\lambda} I_n - J_n^2(\mu)]^{-1},$$

which when iterated yields the expression for Y_1 in (2.12).

The proof of the next result follows that of Lemma 3 and is omitted.

Case III. $\lambda \geqq 0$, and $\mu < 0$ or $\operatorname{Im}(\mu) \neq 0$.

LEMMA 4. *Consider the equation*

$$J_m(\lambda)\bar{Y} - Y\begin{bmatrix} O & I_n \\ J_n(\mu) & O \end{bmatrix} = C$$

where $\lambda \geqq 0$, $\mu < 0$ *or* $\operatorname{Im}(\mu) \neq 0$, *and* C *is an* m*-by-*$2n$ *matrix. Partition* Y *and* C *conformally into* m*-by-*n *submatrices* $Y = [Y_1, Y_2]$ *and* $C = [C_1, C_2]$. *Then there is a unique solution of the form*

$$[J_m(\lambda)\bar{Y}_2 - C_2, Y_2] \quad \text{with } Y_2 = \Phi_R[J_m^2(\lambda) - \mu I_m, C_1 + J_m(\lambda)\bar{C}_2, J_n]. \tag{2.14}$$

We now come to the last case, which has four subcases.

Case IV. $\lambda < 0$ or $\operatorname{Im}(\lambda) \neq 0$, and $\mu < 0$ or $\operatorname{Im}(\mu) \neq 0$.

LEMMA 5. *Consider the equation*

$$\begin{bmatrix} O & I_m \\ J_m(\lambda) & O \end{bmatrix} \bar{Y} - Y \begin{bmatrix} O & I_n \\ J_n(\mu) & O \end{bmatrix} = C$$

where $\lambda < 0$ *or* $\operatorname{Im}(\lambda) \neq 0$, $\mu < 0$, *or* $\operatorname{Im}(\mu) \neq 0$, *and* C *is a* $2m$*-by-*$2n$ *matrix. Partition* Y *and* C *into* m*-by-*n *submatrices*

$$Y = \begin{bmatrix} Y_1 & Y_2 \\ Y_3 & Y_4 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} C_1 & C_2 \\ C_3 & C_4 \end{bmatrix}.$$

Then a solution, if any, has the form

$$Y = \begin{bmatrix} Y_1 & Y_2 \\ \bar{Y}_2 J_n(\bar{\mu}) + \bar{C}_1 & \bar{Y}_1 + \bar{C}_2 \end{bmatrix}. \tag{2.15}$$

The blocks Y_1 *and* Y_2 *are given below, where* $W_1 = \bar{C}_3 + C_2 J_n(\bar{\mu})$ *and* $W_2 = C_1 + \bar{C}_4$.

(i) *If* $\lambda \neq \mu$ *and* $\bar{\lambda} \neq \mu$, *then the solution* (2.15) *is unique with*

$$Y_1 = -\Phi_L[J_m, W_1, J_n(\bar{\mu} - \bar{\lambda})] = \Phi_R[J_m(\bar{\lambda} - \bar{\mu}), W_1, J_n], \tag{2.16}$$

$$Y_2 = -\Phi_L[J_m, W_2, J_n(\mu - \bar{\lambda})] = \Phi_R[J_m(\bar{\lambda} - \mu), W_2, J_n]. \tag{2.17}$$

(ii) *If* $\lambda = \mu < 0$, *then a solution exists if and only if*

$$\Omega_L(W_1)e_1 = O \quad \text{for } m \leqq n \quad \text{or} \quad e_m^T \Omega_R(W_1) = O \quad \text{for } m \geqq n, \tag{2.18}$$

$$\Omega_L(W_2)e_1 = O \quad \text{for } m \leqq n \quad \text{or} \quad e_m^T \Omega_R(W_2) = O \quad \text{for } m \geqq n. \tag{2.19}$$

In this case a particular solution Y_p *is given by* (2.15) *with*

$$Y_1 = -\Omega_L(W_1)J_n^T \quad \text{for } m \leqq n \quad \text{or} \quad Y_1 = J_m^T \Omega_R(W_1) \quad \text{for } m \geqq n, \tag{2.20}$$

$$Y_2 = -\Omega_L(W_2)J_n^T \quad \text{for } m \leqq n \quad \text{or} \quad Y_2 = J_m^T \Omega_R(W_2) \quad \text{for } m \geqq n. \tag{2.21}$$

The general solution is then given by

$$Y = Y_p + \begin{bmatrix} U & V \\ J_m(\lambda)\bar{V} & \bar{U} \end{bmatrix}$$

where U *and* V *are arbitrary* m*-by-*n *upper Toeplitz matrices.*

(iii) *If* $\lambda = \mu \neq \bar{\mu}$, *then a solution exists if and only if the consistency condition* (2.18) *is satisfied. In this case a particular solution* Y_p *has the form* (2.15), *where* Y_1 *is given by* (2.20) *and* Y_2 *is given by* (2.17). *The general solution is then given by*

$$Y = Y_p + \begin{bmatrix} U & O \\ O & \bar{U} \end{bmatrix}$$

where U *is an arbitrary* m*-by-*n *upper-Toeplitz matrix.*

(iv) *If* $\lambda \neq \bar{\lambda} = \mu$, *then a solution exists if and only if the consistency condition* (2.19) *is satisfied. In this case a particular solution* Y_p *has the form* (2.15), *where* Y_1 *is given by* (2.16) *and* Y_2 *is given by* (2.21). *The general solution is then given by*

$$Y = Y_p + \begin{bmatrix} O & V \\ J_m(\lambda)\bar{V} & O \end{bmatrix}$$

where V *is an arbitrary* m*-by-*n *upper-Toeplitz matrix.*

Proof. By writing the equation in terms of the partition submatrices we obtain

$$\begin{aligned} &\text{(a)} \quad \bar{Y}_3 - Y_2 J_n(\mu) = C_1 \quad \text{or} \quad Y_3 = \bar{Y}_2 J_n(\bar{\mu}) + \bar{C}_1, \\ &\text{(b)} \quad \bar{Y}_4 - Y_1 = C_2 \quad \text{or} \quad Y_4 = \bar{Y}_1 + \bar{C}_2, \\ &\text{(c)} \quad J_m(\lambda)\bar{Y}_1 - Y_4 J_n(\mu) = C_3, \\ &\text{(d)} \quad J_m(\lambda)\bar{Y}_2 - Y_3 = C_4. \end{aligned} \tag{2.22}$$

Now (2.22a) and (2.22b) immediately imply that Y has the form of (2.15). Substituting for Y_4 in (2.22c) and for Y_3 in (2.22d) we obtain

$$\text{(2.23)} \quad \begin{aligned} &\text{(a)} \quad J_m(\bar{\lambda})Y_1 - Y_1 J_n(\bar{\mu}) = \bar{C}_3 + C_2 J_n(\bar{\mu}) = W_1, \\ &\text{(b)} \quad J_m(\bar{\lambda})Y_2 - Y_2 J_n(\mu) = C_1 + \bar{C}_4 = W_2. \end{aligned}$$

We now directly apply Lemma 1 to the last two equations to yield the desired results. The complementary solutions are given by Theorem 2(4), 2(6) and Lemma 6(4) of [2].

In summary of the results of this section we have the following theorem.

THEOREM 1. *For square complex matrices A and B, where A is m-by-m and B is n-by-n, let $K = PA\bar{P}^{-1}$ and $L = QB\bar{Q}^{-1}$ be the concanonical forms of A and B, respectively. Denote the diagonal blocks of K and L by K_s and L_t, respectively. For a given m-by-n complex matrix C, partition $PC\bar{Q}^{-1} = [C_{st}]$ with row blocks conformal to K and column blocks conformal to L. Then $A\bar{X} - XB = C$ is consistent if and only if $K_s\bar{Y}_{st} - Y_{st}L_t = C_{st}$ is consistent for all s and t. If $A\bar{X} - XB = C$ is consistent, then X is a solution of this equation if and only if $X = P^{-1}[Y_{st}]Q$, where the Y_{st} are solutions of $K_s\bar{Y}_{st} - Y_{st}L_t = C_{st}$ as given in Lemmas 2–5.*

Note that if λ_s and μ_t are the conspectral values associated with the blocks K_s and L_t, then $K_s\bar{Y}_{st} - Y_{st}L_t = C_{st}$ is immediately known to be consistent if $C_{st} = O$ or if $\lambda_s \neq \mu_t$ and $\bar{\lambda}_s \neq \mu_t$; otherwise consistency may be checked by using the appropriate condition (2.5), (2.7), (2.18), or (2.19) as given in Lemmas 2 and 5.

When $B = A$ and $Q = P$, Theorem 1 may be applied directly to the equation $A\bar{X} - XA = C$. Furthermore, X is a solution to $A\bar{X} + XA = C$ if and only if iX is a solution to $A(\bar{iX}) - (iX)A = -iC$. Next we note that the blocks of $\bar{K} = \bar{P}\bar{A}P^{-1}$ have the same form as the blocks considered in Lemmas 2–5 and $A\bar{X} \pm X\bar{A} = C$ is equivalent to $K(\overline{PXP^{-1}}) \pm (PXP^{-1})\bar{K} = PCP^{-1}$. Thus Lemmas 2–5 are sufficient to characterize solutions of the equations $A\bar{X} \pm XA = C$ and $A\bar{X} \pm X\bar{A} = C$.

3. The equations $A\bar{X} \pm XA^T = C$ and $A\bar{X} \pm XA^* = C$. As an application of the results of § 2 let us consider the special cases of $A\bar{X} \pm XB = C$, where $B = A^T$ or $B = A^*$. In order to apply Lemmas 2–5 we will need properties of the reverse identity or "flip" matrix

$$F = \begin{bmatrix} & & & 1 \\ & O & \iddots & \\ & 1 & & O \\ 1 & & & \end{bmatrix}$$

for which $F = \bar{F} = F^T = F^{-1}$ and

$$FJ_n(\mu)F = J_n^T(\mu). \tag{3.1}$$

Let $K = PA\bar{P}^{-1}$ be the concanonical form of A with diagonal blocks K_s, and let G be the block diagonal matrix with diagonal blocks G_s conformal to those of K such that $G_s = F$ if $K_s = J_k(\lambda)$, or

$$G_s = \begin{bmatrix} O & F \\ F & O \end{bmatrix} \quad \text{if } K_s = \begin{bmatrix} O & I \\ J_k(\lambda) & O \end{bmatrix}.$$

Thus $G = \bar{G} = G^T = G^{-1}$ and $GK = K^TG$. We will say that G is the block flip matrix determined by K. Since $K^T = P^{*-1}A^TP^T$, premultiplication by P and postmultiplication by P^T show that

$$\begin{aligned} A\bar{X} \pm XA^T = C &\Leftrightarrow K\bar{Z} \pm ZK^T = PCP^T \quad \text{with } Z = PXP^* \\ &\Leftrightarrow K\bar{Y} \pm YK = PCP^TG \quad \text{with } Y = ZG = PXP^*G. \end{aligned} \tag{3.2}$$

Similarly $K^* = P^{T-1}A^*P^*$ so that

$$
\begin{aligned}
A\bar{X} \pm XA^* = C &\Leftrightarrow K\bar{Z} \pm ZK^* = PCP^* \quad \text{with } Z = PXP^T \\
&\Leftrightarrow K\bar{Y} \pm Y\bar{K} = PCP^*G \quad \text{with } Y = ZG = PXP^TG.
\end{aligned} \tag{3.3}
$$

Thus we obtain the following theorem.

THEOREM 2. *For square complex matrices A and C, let $K = PA\bar{P}^{-1}$ be the concanonical form of A. Let G be the block flip matrix determined by K, and denote the diagonal blocks of K and G by K_s and G_s, respectively.*

(i) *If $PCP^T = [C_{st}]$ is partitioned conformally to K, then $A\bar{X} \pm XA^T = C$ is consistent if and only if $K_s\bar{Y}_{st} \pm Y_{st}K_t = C_{st}G_t$ is consistent for all s and t. In which case X is a solution of $A\bar{X} \pm XA^T = C$ if and only if $X = P^{-1}[Y_{st}]GP^{*-1}$, where the Y_{st} are solutions of $K_s\bar{Y}_{st} \pm Y_{st}K_t = C_{st}G_t$ as given in § 2.*

(ii) *If $PCP^* = [C_{st}]$ is partitioned conformally to K, then $A\bar{X} \pm XA^* = C$ is consistent if and only if $K_s\bar{Y}_{st} \pm Y_{st}\bar{K}_t = C_{st}G_t$ is consistent for all s and t. In which case X is a solution of $A\bar{X} \pm XA^* = C$ if and only if $X = P^{-1}[Y_{st}]GP^{T-1}$, where the Y_{st} are solutions of $K_s\bar{Y}_{st} \pm Y_{st}\bar{K}_t = C_{st}G_t$ as given in § 2.*

Note that if X is Hermitian and C is defined by $C = A\bar{X} - XA^T$, then $C^T = XA^T - AX^T = XA^T - A\bar{X} = -C$ so C is skew-symmetric. A converse of this also holds in that if C is skew-symmetric and $A\bar{X} - XA^T = C$ is consistent, then the equation has a Hermitian solution. To obtain this converse, let X be some solution to the equation. It is then easy to check that $\frac{1}{2}[X + X^*]$ is a Hermitian solution. Using (3.2) and (3.3) we can obtain a little more.

THEOREM 3.

(i) *For the equation $A\bar{X} - XA^T = C$ the following are equivalent:*
  (a) *$C^T = -C$ and the equation is consistent,*
  (b) *The equation has a Hermitian solution,*
  (c) *The equation has a nonsingular Hermitian solution.*

(ii) *For the equation $A\bar{X} - XA^* = C$ the following are equivalent:*
  (a) *$C^* = -C$ and the equation is consistent,*
  (b) *The equation has a symmetric solution,*
  (c) *The equation has a nonsingular symmetric solution.*

(iii) *For the equation $A\bar{X} + XA^T = C$ the following are equivalent:*
  (a) *$C^T = -C$ and the equation is consistent,*
  (b) *The equation has a skew-Hermitian solution,*
  (c) *The equation has a nonsingular skew-Hermitian solution.*

(iv) *For the equation $A\bar{X} + XA^* = C$ the following are equivalent:*
  (a) *$C^* = C$ and the equation is consistent,*
  (b) *The equation has a symmetric solution,*
  (c) *The equation has a nonsingular symmetric solution.*

Proof. In each part it is clear that (c) implies (b) which implies (a). All we must do is show that (a) implies (c) in each part. For part (i) let X_p be a particular solution so that $\frac{1}{2}[X_p + X_p^*]$ is a particular Hermitian solution as noted above. By (3.2), $A\bar{X} - XA^T = O$ is equivalent to $K\bar{Z} - ZK^T = O$, where $Z = PXP^*$. Thus we may take $Z_c = G$ so that $X_c = P^{-1}GP^{-1*}$ is a nonsingular Hermitian solution to the complementary equation. Hence $\frac{1}{2}[X_p + X_p^*] + \alpha X_c$ must be a nonsingular Hermitian solution for some real value of the scalar α. Part (ii) may be obtained in a similar manner. To be specific, if X_p is a particular solution to the equation of part (ii) and C is skew-Hermitian then $\frac{1}{2}[X_p + X_p^T] + \alpha P^{-1}KGP^{-1T}$ is a nonsingular symmetric solution for some value of the scalar α. Parts (iii) and (iv) follow from (i) and (ii), since X is a nonsingular skew-Hermitian solution of $A\bar{X} + XA^T = C$ if and only if iX is a nonsingular Hermitian solution

of $A(\bar{i}\bar{X}) - (iX)A^T = -iC$, and X is a nonsingular symmetric solution of $A\bar{X} + XA^* = C$ if and only if iX is a nonsingular symmetric solution of $A(\bar{i}\bar{X}) - (iX)A^* = -iC$.

Hong and Horn [5] have shown that every square matrix is consimilar to its transpose and to its conjugate transpose. By taking $C = O$ in Theorem 3 we obtain the following corollary.

COROLLARY 1. *If A is a square matrix, then:*

(i) *A is consimilar to A^T via a Hermitian matrix,*

(ii) *A is consimilar to A^* via a symmetric matrix,*

(iii) *A is consimilar to $-A^T$ via a skew-Hermitian matrix,*

(iv) *A is consimilar to $-A^*$ via a symmetric matrix.*

Corollary 1 may also be obtained by an argument similar to that given in [7, p. 172] to show that A is similar to A^* if and only if A is similar to A^* via a Hermitian similarity transformation. Another interpretation of the results of Corollary 1 is given by the following corollary.

COROLLARY 2. *Any square matrix A can be written as a product $A = SH$ or $A = HS$ where S is symmetric and H is Hermitian or skew-Hermitian. When A is singular exactly one of S or H can be chosen to be nonsingular (when A is nonsingular both S and H are nonsingular).*

Proof. This corollary describes eight possibly different factorizations of a matrix A. One of these follows from Corollary 1(i), which provides a nonsingular Hermitian matrix H such that $AH = \bar{H}A^T$. Thus $A = (\bar{H}A^T)H^{-1}$, where H^{-1} is nonsingular Hermitian and $(\bar{H}A^T)^T = AH^* = AH = \bar{H}A^T$ is symmetric. Three other factorizations are similarly obtained from Corollary 1(ii)–(iv). The other four factorizations are obtained by applying parts (i) and (iii) of Corollary 1 to A^T, and parts (ii) and (iv) to A^*.

Elementary calculations similar to those preceding Theorem 3 show that:

(i) $A\bar{X} - XA^T = C$ is consistent and $C = C^T$
$\Leftrightarrow A\bar{X} - XA^T = C$ has a skew-Hermitian solution;

(ii) $A\bar{X} - XA^* = C$ is consistent and $C = C^*$
$\Leftrightarrow A\bar{X} - XA^* = C$ has a skew-symmetric solution;

(iii) $A\bar{X} + XA^T = C$ is consistent and $C = C^T$
$\Leftrightarrow A\bar{X} + XA^T = C$ has a Hermitian solution;

(iv) $A\bar{X} + XA^* = C$ is consistent and $C = -C^*$
$\Leftrightarrow A\bar{X} + XA^* = C$ has a skew-symmetric solution.

However, we cannot necessarily claim that these special solutions are nonsingular as in Theorem 3. The reason for this may be seen by taking $C = O$ in Theorem 4.

THEOREM 4. *Suppose that $A\bar{A}$ is nonderogatory and nonsingular.*

(i) *If $C^T = -C$ and $A\bar{X} - XA^T = C$ is consistent, then every solution is Hermitian.*

(ii) *If $C^* = -C$ and $A\bar{X} - XA^* = C$ is consistent, then every solution is symmetric.*

(iii) *If $C^T = -C$ and $A\bar{X} + XA^T = C$ is consistent, then every solution is skew-Hermitian.*

(iv) *If $C^* = C$ and $A\bar{X} + XA^* = C$ is consistent, then every solution is symmetric.*

Proof. The nonsingularity of $A\bar{A}$ implies that 0 is not a coneigenvalue of A. Also if

$$\begin{bmatrix} O & I \\ J(\lambda) & O \end{bmatrix}, \qquad \lambda < 0$$

is a block of the concanonical form of A, then (see [2, § 2])

$$\begin{bmatrix} J(\lambda) & O \\ O & J(\lambda) \end{bmatrix}$$

is contained in the Jordan form of $A\bar{A}$. Since $A\bar{A}$ is nonderogatory, there can be no negative coneigenvalues of A. Thus if λ_s and λ_t are the conspectral values associated with the blocks K_s and K_t of the concanonical form of A, where $s \neq t$, then $\lambda_s \neq \lambda_t$, $\lambda_s > 0$ or $\operatorname{Im}(\lambda_s) > 0$, and $\lambda_t > 0$ or $\operatorname{Im}(\lambda_t) > 0$. We use the notation of Theorem 2 so that for part (i) $A\bar{X} - XA^T = C$ is equivalent to $X = P^{-1}[Z_{st}]P^{-1}{}^*$ with $Z_{st} = Y_{st}G_t$, where $K_s\bar{Y}_{st} - Y_{st}K_t = C_{st}G_t$. Thus we need to show that $Z = [Z_{st}]$ is Hermitian or $Z_{st} = Z_{ts}^*$ for all s, t. Now $[C_{st}] = PCP^T$ is skew-symmetric so that $C_{st}^T = -C_{ts}$ for all s, t and $Y_{st}^*K_s^T - K_t^TY_{st}^T = -G_tC_{ts}$. Hence $K_t(G_tY_{st}^*G_s) - (G_tY_{st}^*G_s)K_s = C_{ts}G_s$ so that $G_tY_{st}^*G_s$ is a solution to the equation for Y_{ts}. For $s \neq t$ this solution is unique by Lemmas 2(i), 3, 4, or 5(i), and thus $Y_{ts} = G_tY_{st}^*G_s$ or $Z_{ts} = Y_{ts}G_s = (Y_{st}G_t)^* = Z_{st}^*$. Also for $s = t$, $Y_{ss} - G_sY_{ss}^*G_s$ is a solution to the homogeneous equation $K_s\bar{Y} - YK_s = O$. If $K_s = J_m(\lambda_s)$, $\lambda_s > 0$, then by Lemma 2(ii), $Y_{ss}G_s - G_sY_{ss}^* = [Y_{ss} - G_sY_{ss}^*G_s]\,G_s = VG_s$, where V is an m-by-m real upper Toeplitz matrix. Now V is a real linear combination of powers of J_m so that $VG_s = VF = FV^T = G_sV^* = (VG_s)^* = [Y_{ss}G_s - G_sY_{ss}^*]^* = -VG_s$. Hence $VG_s = O$ and $Z_{ss} = Y_{ss}G_s = G_sY_{ss}^* = Z_{ss}^*$. Similarly, if

$$K_s = \begin{bmatrix} O & I_m \\ J_m(\lambda_s) & O \end{bmatrix}, \qquad \operatorname{Im}(\lambda_s) > 0,$$

then by Lemma 5(iii)

$$Y_{ss} - G_sY_{ss}^*G_s = \begin{bmatrix} V & O \\ O & \bar{V} \end{bmatrix}$$

where V is an m-by-m upper Toeplitz matrix. Thus

$$\begin{aligned} G_sY_{ss}^*G_s - Y_{ss} &= G_s[Y_{ss} - G_sY_{ss}^*G_s]^*G_s \\ &= \begin{bmatrix} O & F \\ F & O \end{bmatrix}\begin{bmatrix} V^* & O \\ O & V^T \end{bmatrix}\begin{bmatrix} O & F \\ F & O \end{bmatrix} = \begin{bmatrix} V & O \\ O & \bar{V} \end{bmatrix} = Y_{ss} - G_sY_{ss}^*G_s. \end{aligned}$$

As before $Z_{ss} = Y_{ss}G_s = G_sY_{ss}^* = Z_{ss}^*$, which completes part (i).

Part (ii) is similar to part (i) except that we also need the equality $J_m(\lambda)V = VJ_m(\lambda)$ which holds for any m-by-m upper Toeplitz matrix. Parts (iii) and (iv) follow from parts (i) and (ii) as in the proof of Theorem 3.

Comment. Suppose that $A\bar{X} - XA^T = C$ is consistent, C is skew-symmetric, and A is any square matrix. One approach to obtaining a Hermitian solution to this equation could be: find Z_{st} such that $K_s\bar{Z}_{st} - Z_{st}K_t^T = C_{st}$ for all s, t, where $PCP^T = [C_{st}]$, form $X = P^{-1}[Z_{st}]P^{-1}{}^*$, and then form $\frac{1}{2}[X + X^*]$. However, the proof of Theorem 4 suggests an alternative method, namely: (1) find Z_{st} such that $K_s\bar{Z}_{st} - Z_{st}K_t^T = C_{st}$ for all $s \geqq t$; (2) set $Z_{st} = Z_{ts}^*$ for all $s < t$; (3) replace Z_{ss} by $\frac{1}{2}[Z_{ss} + Z_{ss}^*]$ for $\lambda_s \leqq 0$; (4) form $X = P^{-1}[Z_{st}]P^{-1}{}^*$. Clearly X is Hermitian if and only if $[Z_{st}]$ is Hermitian, so this alternative method is justified by the observations in the proof of Theorem 4 that when C is skew-symmetric Z_{ts}^* is a solution to the equation for Z_{st}, and $Z_{ss} = Z_{ss}^*$ for $\lambda_s > 0$ or $\operatorname{Im}(\lambda_s) > 0$. Thus steps (1) and (2) provide a matrix $[Z_{st}]$ which is Hermitian for all off-diagonal blocks and for all diagonal blocks with $\lambda_s > 0$ or $\operatorname{Im}(\lambda_s) > 0$. That is, step (3) is only necessary for diagonal blocks with $\lambda_s \leqq 0$.

The blocks Z_{st} required for this approach may be obtained from Lemmas 2–5 since they are determined by $Z_{st} = Y_{st}G_t$, where $K_s\bar{Y}_{st} - Y_{st}K_t = C_{st}G_t$.

In relation to the matrix $Z_{ss} + Z_{ss}^* = Y_{ss}G_s + (Y_{ss}G_s)^*$ in step (3), it is interesting to note that the matrices $Y_{ss}G_s$ and $(Y_{ss}G_s)^*$ are related to the two forms of the solutions given by Lemma 2(iii) and Lemma 5(ii). We exhibit this relationship below, where we

use the properties $[\Gamma_L(W)]^T = \Gamma_L(W^T)$, $[\Gamma_R(W)]^T = \Gamma_R(W^T)$, and $F\Gamma_L(W)F = \Gamma_R(FWF)$ of the Γ notation, and similar properties of the Ω notation.

First consider the case $K_s = J_m(0)$, where $\lambda_s = 0$. Since $m = n$ in $K_s\bar{Y}_{ss} - Y_{ss}K_s = C_{ss}G_s$, (2.8) yields two solutions $Y_1 = -\Gamma_L(C_{ss}G_s)J_m^T$ and $Y_2 = J_m^T\Gamma_R(\bar{C}_{ss}G_s)$. Since $C_{ss} = -C_{ss}^T$ we obtain

$$G_sY_1^*G_s = FJ_m\Gamma_L(-C_{ss}F)^*F = FJ_m\Gamma_L(-F\bar{C}_{ss}^T)F = FJ_m\Gamma_L(F\bar{C}_{ss})F = J_m^T\Gamma_R(\bar{C}_{ss}G_s) = Y_2.$$

Hence Y_1G_s and $(Y_1G_s)^* = Y_2G_s$ are both solutions to the equation for Z_{ss}. Thus $\frac{1}{2}[Y_1 + Y_2]G_s$ is a Hermitian solution to the equation for Z_{ss}.

We next consider the case

$$K_s = \begin{bmatrix} O & I_m \\ J_m(\lambda_s) & O \end{bmatrix} \quad \text{with } \lambda_s < 0.$$

Partition Y_{ss} and C_{ss} as

$$\begin{bmatrix} Y_1 & Y_2 \\ Y_3 & Y_4 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} C_1 & C_2 \\ C_3 & C_4 \end{bmatrix},$$

respectively, so that $C_1^T = -C_1$, $C_4^T = -C_4$, $C_2^T = -C_3$, and

$$Z_{ss} = Y_{ss}G_s = \begin{bmatrix} Y_2F & Y_1F \\ Y_4F & Y_3F \end{bmatrix}$$

where

$$K_s\bar{Y}_{ss} - Y_{ss}K_s = \begin{bmatrix} C_2F & C_1F \\ C_4F & C_3F \end{bmatrix}.$$

We now show that Z_{ss} is Hermitian whenever Y_2 is chosen such that Y_2F is Hermitian:

(i) If $Y_2F = (Y_2F)^*$, then by (2.22a) and (2.22d), $Y_3 = \bar{Y}_2J_m(\lambda_s) + \bar{C}_2F = J_m(\lambda_s)\bar{Y}_2 - C_3F$ so that

$$(Y_3F)^* = F[J_m^T(\lambda_s)Y_2^T + FC_2^T] = J_m(\lambda_s)(\bar{Y}_2F)^* + C_2^T = J_m(\lambda_s)\bar{Y}_2F - C_3 = Y_3F.$$

(ii) Let $U = Y_1F - (Y_4F)^*$. By (2.22b) $Y_4 = \bar{Y}_1 + \bar{C}_1F$ so that $UF = Y_1 - FY_1^TF - C_1^TF$. We may use (2.23a) which in this case states that $J_m(\lambda_s)Y_1 - Y_1J_m(\lambda_s) = \bar{C}_4F + C_1FJ_m(\lambda_s)$ to compute $J_m(\lambda_s)(ZF) - (ZF)J_m(\lambda_s) = O$. Hence by Lemma 1, UF is an m-by-m upper Toeplitz matrix. Now

$$U = (UF)F = F(UF)^T = U^T = [Y_1F - FY_1^T + C_1]^T = FY_1^T - Y_1F - C_1 = -U.$$

Thus $U = O$ and $Y_1F = (Y_4F)^*$.

We complete the case $\lambda_s < 0$ by showing how to choose Y_2 such that Y_2F is Hermitian. Since $m = n$, (2.21) provides two solutions to (2.23b). By taking half the sum of these two solutions we obtain another solution of (2.23b), namely,

$$Y_2 = \tfrac{1}{2}[J_m^T\Omega_R(W_2) - \Omega_L(W_2)J_m^T],$$

where $W_2 = C_2F + \bar{C}_3F$. Now $FW_2^*F = -W_2$ so that

$$\begin{aligned}(Y_2F)^* &= \tfrac{1}{2}F[\Omega_R(W_2^*)J_m - J_m\Omega_L(W_2^*)] = \tfrac{1}{2}[\Omega_L(FW_2^*F)J_m^T - J_m^T\Omega_R(FW_2^*F)]F \\ &= \tfrac{1}{2}[-\Omega_L(W_2)J_m^T + J_m^T\Omega_R(W_2)]F = Y_2F.\end{aligned}$$

REFERENCES

[1] S. BARNETT AND M. J. C. GOVER, *Some extensions of Hankel and Toeplitz matrices*, Linear and Multilinear Algebra, 14 (1983), pp. 45–65.

[2] J. BEVIS, F. J. HALL, AND R. E. HARTWIG, *Consimilarity and the matrix equation* $A\bar{X} - XB = C$, in Current Trends in Matrix Theory, F. Uhlig and R. Grone, eds., North–Holland, Amsterdam, 1987, pp. 51–64.

[3] R. E. HARTWIG, $AX - XB = C$, *resultants, and generalized inverses*, SIAM J. Appl. Math., 28 (1975), pp. 154–183.

[4] Y. P. HONG, *Consimilarity: theory and applications*, Ph.D. thesis, The John Hopkins Univ., Baltimore, MD, 1985.

[5] Y. P. HONG AND R. A. HORN, *A canonical form for matrices under consimilarity*, Linear Algebra Appl., to appear.

[6] ———, *On the reduction of a matrix to triangular or diagonal form by consimilarity*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 80–88.

[7] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, U.K., 1985.

[8] N. JACOBSON, *Pseudo-linear transformations*, Ann. of Math., 38 (1937), pp. 484–507.

[9] ———, *The Theory of Rings*, American Mathematical Society, Providence, RI, 1943.

[10] D. E. RUTHERFORD, *On the solution of the matrix equation* $AX + XB = C$, Nederl. Akad. Wetensch. Proc. Ser. A, 35 (1932), pp. 53–59.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

007

ANALYSIS AND SOLUTION OF THE NONGENERIC TOTAL LEAST SQUARES PROBLEM*

SABINE VAN HUFFEL†‡ AND JOOS VANDEWALLE†

Abstract. Total least squares (TLS) is one method of solving overdetermined sets of linear equations $AX \approx B$ that is appropriate when there are errors in both the observation matrix B and the data matrix A. Golub and Van Loan (G. H. Golub and C. F. Van Loan, *SIAM J. Numer. Anal.*, 17 (1980), pp. 883–893) introduced this method into the field of numerical analysis and developed an algorithm based on the singular value decomposition. However in some TLS problems, called nongeneric, their algorithm fails to compute a finite TLS solution. This paper generalizes their TLS computations in order to solve these nongeneric TLS problems. The authors describe the properties of those problems and prove that the proposed generalization remains optimal with respect to the TLS criteria for any number of observation vectors in B if additional constraints are imposed. Finally, the TLS computation is summarized in one algorithm which includes the proposed generalization.

Key words. total least squares, singular value decomposition, overdetermined sets of equations, numerical linear algebra

AMS(MOS) subject classifications. 15A18, GSF20

1. Introduction.

Many problems in signal processing, system theory, automatic control and in general engineering, physics, and economics give rise to an *overdetermined* set of linear equations $AX \approx B$ which are usually solved with the linear *least squares* (LS) technique. This technique assumes that all the *errors can be allocated to the observation matrix B.* Unfortunately, this assumption is frequently unrealistic; sampling errors, human errors, modeling errors, and instrument errors may imply inaccuracies on the data matrix A. For those cases a better *more general* fitting technique, *total least squares* (TLS), has been devised to compensate for data errors. The TLS approach is appropriate when independent and equally sized errors occur in all data and amounts to fitting a best subspace to the data. Although studies of the univariate problem, i.e., line fitting, are quite old [1], the multivariate problem has only been analyzed recently. Golub and Van Loan [5] introduced the method in the field of numerical analysis by presenting first a *singular value decomposition* (SVD) analysis of the problem.

The TLS problem can be formulated as follows (R denotes the range).

TLS DEFINITION. Given an overdetermined set of m linear equations in $n \times d$ unknowns

$$AX \approx B, \quad A \in \mathbb{R}^{m \times n}, \quad B \in \mathbb{R}^{m \times d}, \quad X \in \mathbb{R}^{n \times d}, \tag{1}$$

a TLS *solution* is any solution X of the set

$$\hat{A}X = \hat{B} \tag{2}$$

where $\hat{A}$ and $\hat{B}$ are determined such that

$$R(\hat{B}) \subset R(\hat{A}), \tag{3}$$

$$\|[\Delta\hat{A}; \Delta\hat{B}]\|_F = \|[A; B] - [\hat{A}; \hat{B}]\|_F \quad \text{is minimal.} \tag{4}$$

The problem of finding $[\Delta\hat{A}; \Delta\hat{B}]$ such that (3)–(4) are satisfied, is referred to as the TLS *problem.*

* Received by the editors August 17, 1987; accepted for publication October 19, 1987.

† Electronics, Systems, Automation, and Technology (ESAT) Laboratory, Department of Electrical Engineering, K. U. Leuven, Kardinaal Mercierlaan 94, B-3030 Heverlee, Belgium.

‡ This author is a senior research assistant of the Belgian National Fund of Scientific Research (NFWO).

Whenever the TLS solution is not unique, TLS singles out the *minimum norm* solution. It is clear that the TLS solution can be deduced from a basis of the kernel of $[\hat{A}; \hat{B}]$. This computation can be done generically by using the algorithm of Golub and Van Loan [5]. However in some cases, the computed solution becomes *infinite* and does not satisfy condition (4). For those cases, Golub and Van Loan concluded that the TLS problem has no solution. In this paper however, we *generalize the* TLS *computations to all nongeneric* TLS *problems*, i.e., problems in which the algorithm of Golub and Van Loan fails to compute a TLS solution. We describe the properties of those nongeneric problems and prove that the proposed generalization remains optimal with respect to the TLS criteria (3)–(4) in the one-dimensional case ($d = 1$), as well as in the multidimensional case ($d > 1$) under additional constraints.

This paper is organized into five sections. In § 2, the nongeneric TLS problem is formulated. Section 3 then describes the properties and solution of the nongeneric problem and compares the TLS solution of those problems with the LS solution. The one-dimensional case, as well as the multidimensional case, is considered. In § 4, the TLS computations are summarized in one generalized TLS algorithm. Finally, § 5 presents the conclusions.

2. Formulation of the nongeneric TLS problem. Before starting, we introduce the *notation* used throughout this paper:

The superscript T denotes the *transpose* of a vector or matrix.

The *m bij m identity* matrix is denoted by I_m.

X' is the n by d *minimum norm* least squares (LS) *solution* and $\hat{X}$ is the n by d minimum norm total least squares (TLS) solution of (1).

For the *one-dimensional* problem, i.e., $d = 1$, the matrices are replaced by their corresponding vector notation, e.g., the vectors $\mathbf{b}$ and $\mathbf{x}$ are used instead of the matrices B and X in (1).

Denote the *singular value decomposition* (SVD) of A in (1) by

$$
\begin{aligned}
&A = U'\Sigma'V'^T \quad \text{with } U' = [U'_1; U'_2],\ U'_1 = [\mathbf{u}'_1, \cdots, \mathbf{u}'_n],\ U'_2 = [\mathbf{u}'_{n+1}, \cdots, \mathbf{u}'_m],\\
&\mathbf{u}'_i \in \mathbb{R}^m, \qquad U'^T U' = I_m,\\
&V' = [\mathbf{v}'_1, \cdots, \mathbf{v}'_n], \qquad \mathbf{v}'_i \in \mathbb{R}^n, \quad V'^T V' = I_n,\\
&\Sigma' = \operatorname{diag}(\sigma'_1, \cdots, \sigma'_n) \quad \text{and} \quad \sigma'_1 \geqq \cdots \geqq \sigma'_n \geqq 0
\end{aligned} \tag{5}
$$

and denote the SVD of $[A; B]$ in (1) by $[A; B] = U\Sigma V^T$ with $U = [U_1; U_2]$, $U_1 = [\mathbf{u}_1, \cdots, \mathbf{u}_n]$, $U_2 = [\mathbf{u}_{n+1}, \cdots, \mathbf{u}_m]$, $\mathbf{u}_i \in \mathbb{R}^m$,

$$
\begin{aligned}
&U^T U = I_m,\\
&V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{matrix} n \\ d \end{matrix} = [\mathbf{v}_1, \cdots, \mathbf{v}_n, \mathbf{v}_{n+1}, \cdots, \mathbf{v}_{n+d}], \qquad \mathbf{v}_i \in \mathbb{R}^{n+d},\\
&\qquad\quad n \qquad d\\
&V^T V = I_{n+d},\\
&\Sigma = \operatorname{diag}(\sigma_1, \cdots, \sigma_{n+d}) \quad \text{and} \quad \sigma_1 \geqq \cdots \geqq \sigma_{n+d} \geqq 0.
\end{aligned} \tag{6}
$$

Let $V(\sigma_j)$ (respectively, $U(\sigma_j)$) be the right (respectively, left) singular subspace of $[A; B]$ associated with the singular value σ_j, and let $V'(\sigma_j)$ (respectively, $U'(\sigma_j)$) be the right (respectively, left) singular subspace of A associated with the singular value σ_j.

Ker (V_{22}) represents the kernel of V_{22}.

$[A; B']$ is the LS *approximation* of $[A; B]$ with B' the orthogonal projection of B onto the range $R(A)$ of A. The TLS approximation of $[A; B]$ (defined in § 1) is denoted by $[\hat{A}; \hat{B}]$.

The *correction* matrix $\Delta B' = B - B'$ is then the LS approximation effort and the correction matrix $[\Delta\hat{A}; \Delta\hat{B}] = [A - \hat{A}; B - \hat{B}]$ is the TLS approximation effort in order to obtain a solution of (1).

Using this notation, we call the problem *generic* if, for $\sigma_{n-p} > \sigma_{n-p+1} = \cdots = \sigma_{n+1}$ with $p \geqq 0$, the submatrix

$$V_\gamma = \begin{bmatrix} v_{n+1,n-p+1} & \cdots & v_{n+1,n+d} \\ \vdots & & \vdots \\ v_{n+d,n-p+1} & \cdots & v_{n+d,n+d} \end{bmatrix} \quad \text{with } v_{j,i} \text{ the } j\text{th component of } \mathbf{v}_i \tag{7}$$

of V in (6) has full rank d. If $\sigma_n > \sigma_{n+1}$, this means that V_{22} is nonsingular (or $v_{n+1,n+1} \neq 0$ if $d = 1$). The TLS solution of generic problems can be computed with the algorithm of Golub and Van Loan [5] and is given by

$$\hat{X} = -Z\Gamma^{-1} \tag{8}$$

where Z, Γ are obtained by postmultiplying V_γ in (7) with an orthogonal matrix Q such that

$$V_\gamma Q = \begin{array}{c} \left[\begin{array}{ccc:c} & & & \\ & \text{\Large *} & & Z \\ & & & \\ \hdashline 0 & \cdots & 0 & \Gamma \end{array}\right] \begin{array}{l} n \\ \\ \\ d \end{array} \\ \begin{array}{ccc:c} p & & & d \end{array} \end{array} . \tag{9}$$

For generic one-dimensional TLS problems, i.e., $d = 1$, the generic TLS solution (8) reduces to a simple scaling of the last column vector $\begin{bmatrix} \mathbf{z} \\ \gamma \end{bmatrix}$ in $V_\gamma Q$:

$$\hat{\mathbf{x}} = -\mathbf{z}/\gamma. \tag{10}$$

This uniquely determined TLS solution has indeed the minimal norm $\|\hat{X}\|_2$ and $\|\hat{X}\|_F$ as proven by Golub and Van Loan [6, p. 422] for the one-dimensional case and by Van Huffel [8, p. 30] for the multidimensional case.

Whenever V_γ is singular (or $\gamma = 0$, if $d = 1$), the problem is called *nongeneric*. In this case, the solutions (8) and (10) become infinite. This happens when $\sigma'_n \leqq \sigma_{n+1}$, as shown in the next theorem.

THEOREM 2.1. *Let* (6) *be the* SVD *of* $[A; B]$ *and the* SVD *of* A *be given by* (5):

$$V_{22} \textit{ singular} \Rightarrow \sigma_{n+d} \leqq \sigma'_n \leqq \sigma_{n+1}. \tag{11}$$

Proof. See [8, p. 42] for the proof. □

Observe that the case V_{22} *singular* only happens when $\sigma'_n \leqq \sigma_{n+1}$, i.e., the length of the projection σ'_n of all columns $\mathbf{a}_i$ of A onto its lowest singular vector is smaller than the length of the projection of all columns of $[A; B]$ onto its $(n + 1)$th singular vector, associated with σ_{n+1}. This is the case if A is (*nearly*) *rank-deficient*, i.e., $\sigma'_n \approx 0$, or when the set of equations (1) (or at least one subset $A\mathbf{x}_i \approx \mathbf{b}_i$) is highly incompatible.

Contrary to our algebraic approach, Gleser [3] used a statistical approach to prove under which conditions V_{22} is nonsingular. In his proofs we can find the following very interesting property (the superscript m on a sample quantity indicates that the quantity is calculated from the first m rows of $[A; B]$ in (1)).

THEOREM 2.2. *Let* (6) *be the* SVD *of* $[A; B]$ *and* $[A; B] = [A_0; B_0] + [\Delta A; \Delta B]$. *Assume that the rows of the error matrix* $[\Delta A; \Delta B]$ *are independently and identically distributed with zero mean and common covariance matrix* $\sigma_v^2 I_{n+d}$ (σ_v *unknown scalar*). *Assume further that* $\lim_{m\to\infty} (1/m) A_0^T A_0$ *exists and is positive definite*; *then*

$$\exists m_0 \quad \forall m \geqq m_0: \quad V_{11}^{(m)} \text{ and } V_{22}^{(m)} \text{ are nonsingular.}$$

Proof. See [3, p. 35] for the proof. □

This means that the generic TLS solution $\hat{X}^{(m)} = -V_{12}^{(m)} V_{22}^{(m)-1}$ exists for all $m \geqq m_0$. Hence, the property guarantees that, if the assumptions hold, a TLS solution can always be computed in the generic sense when the set of equations (1) is sufficiently overdetermined; in other words, *a nongeneric* TLS *problem can always be made generic* by adding more equations to (1). This is the case if the data $[A; B]$ are observations of an exact but unobservable relation $A_0 X = B_0$ with A_0 of full rank, and the observation errors $[\Delta A; \Delta B]$ are statistically independent and equally sized (same variance).

Those statistical results agree with our algebraic approach: nongeneric TLS problems occur whenever A_0 is (nearly) rank-deficient or when the set $A_0 X \approx B_0$ is highly incompatible. Although "exact" nongeneric TLS problems seldom occur, close-to-nongeneric TLS problems are not uncommon. Moreover, from a numerical point of view, it is very interesting to investigate their properties and generalize the TLS computations to solve these problems in the TLS sense. Indeed, as shown in Theorems 3.1 and 3.4 of § 3, the TLS problem (1) becomes close to nongeneric when σ'_n approaches σ_{n+1}. In those cases, the generic TLS solution can still be computed but is unstable and becomes even very sensitive to data errors when $\sigma'_n - \sigma_{n+1}$ is very close to zero [5]. Identifying the problem as nongeneric, and computing the nongeneric TLS solution, stabilizes the TLS solution and makes it rather insensitive to data errors [8, p. 103].

3. Properties and solution of the nongeneric TLS problem.

3.1. The one-dimensional case. Let us first assume that $\sigma_n \neq \sigma_{n+1}$. In this case the generic TLS solution $\hat{\mathbf{x}}$ in (10) is given by the only one right singular vector of $[A; \mathbf{b}]$, associated with its smallest singular value σ_{n+1}:

$$[\hat{\mathbf{x}}^T; -1]^T = -\mathbf{v}_{n+1}/v_{n+1,n+1} \tag{12}$$

and the TLS approximation is:

$$[\hat{A}; \hat{b}] = U \hat{\Sigma} V^T \quad \text{and} \quad \hat{\Sigma} = \operatorname{diag}(\sigma_1, \cdots, \sigma_n, 0). \tag{13}$$

If $v_{n+1,n+1} = 0$ then the approximation $[\hat{A}; \hat{\mathbf{b}}]$ proposed in (13) does *not* satisfy the TLS condition (3). For those cases, Golub and Van Loan [5] argue that the TLS problem has no solution because (12) becomes infinite. In this section, however, we claim that a TLS approximation $[\hat{A}; \hat{\mathbf{b}}]$, satisfying both (3) and (4) under additional constraints, still exists and can be determined by making the next larger singular value σ_n of $[A; \mathbf{b}]$ in (13) zero. This is equivalent to the case that TLS searches for a solution in a lower-dimensional subspace $[\hat{\hat{A}}; \hat{\hat{\mathbf{b}}}]$, obtained by making one more σ_i in (13) zero, i.e., $\sigma_n = \sigma_{n+1} = 0$. Using Theorem 3.1 (see further) we can indeed prove that under additional constraints this solution is still optimal with respect to (4) and so (3) is satisfied. Moreover it is proven that LS searches its solution in a subspace of the same dimensionality, i.e., there does not exist any LS solution in a subspace of higher dimensionality than that of the TLS approximation. The need for the extension can be best motivated with an example (see Fig. 1).

Example 3.1. Consider the set of three equations in two unknowns:

$$\begin{bmatrix} 2\sqrt{3/2} & 2\sqrt{3/2} \\ \dfrac{1}{2\sqrt{2}} & -\dfrac{1}{2\sqrt{2}} \\ \dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ -\sqrt{3} \end{bmatrix}. \tag{14}$$

Taking the SVD of $[A; \mathbf{b}]$ we obtain

$$U\Sigma V^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & .5 \end{bmatrix} \begin{bmatrix} \dfrac{\sqrt{3}}{2\sqrt{2}} & \dfrac{\sqrt{3}}{2\sqrt{2}} & \dfrac{1}{2} \\ \dfrac{1}{2\sqrt{2}} & \dfrac{1}{2\sqrt{2}} & \dfrac{-\sqrt{3}}{2} \\ \dfrac{1}{\sqrt{2}} & -\dfrac{1}{\sqrt{2}} & 0 \end{bmatrix}. \tag{15}$$

Note that $v_{n+1,n+1} = v_{3,3} = 0$. Since $\mathbf{b} = \sigma_1 v_{3,1}\mathbf{u}_1 + \sigma_2 v_{3,2}\mathbf{u}_2 + 0 \cdot \mathbf{u}_3$ this implies that $\mathbf{b} \perp \mathbf{u}_3$. By taking the rank-$n$ approximation (14), we obtain the approximating vectors

$$\hat{\mathbf{a}}_1 = \hat{\mathbf{a}}_2 = 2\sqrt{\frac{3}{2}}\mathbf{u}_1 + \frac{1}{\sqrt{2}}\mathbf{u}_2, \qquad \hat{\mathbf{b}} = \mathbf{b}. \tag{16}$$

Since $\hat{\mathbf{b}} \notin R(\hat{A})$, (3) is not satisfied. Hence, $[\hat{A}; \hat{\mathbf{b}}]$ as defined in (16) is not a valid TLS approximation and cannot produce a TLS solution in the generic sense. Therefore, Golub and Van Loan [5] argue that no TLS solution exists. However, conceptually there is no

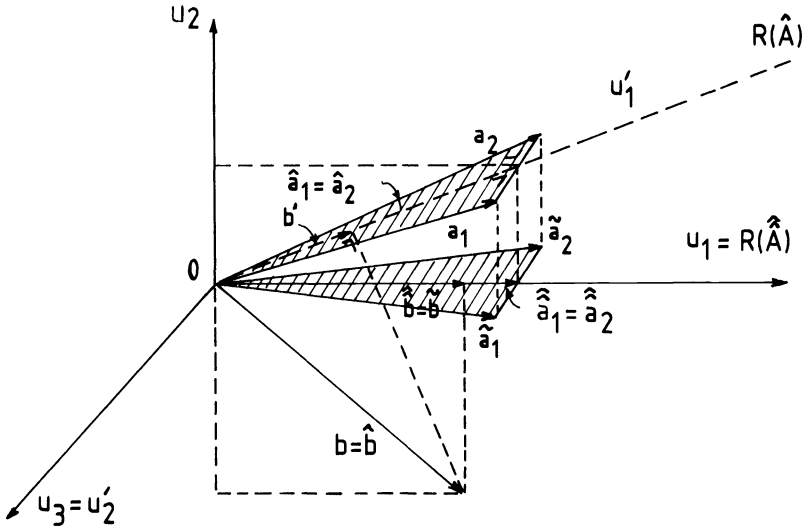


FIG. 1. *Geometric representation of Example* 3.1 *with three equations in two unknowns, characterized by* $v_{3,3} = 0$. *This implies that* $\mathbf{b} \perp \mathbf{u}_3$. $\hat{\mathbf{a}}_1$, $\hat{\mathbf{a}}_2$ *and* $\tilde{\mathbf{b}}$ *are rank-two approximations of* A *and* $\mathbf{b}$ *with* $\tilde{\mathbf{b}} \notin R(\hat{A})$, $\hat{\tilde{\mathbf{a}}}_1$, $\hat{\tilde{\mathbf{a}}}_2$ *and* $\hat{\tilde{\mathbf{b}}}$ *are rank-one approximations of* A *and* $\mathbf{b}$ *with* $\hat{\tilde{\mathbf{b}}} \in R(\hat{\tilde{A}})$ *and* $\tilde{\mathbf{a}}_1$, $\tilde{\mathbf{a}}_2$, *and* $\tilde{\mathbf{b}}$ *are the* TLS *approximations of* A *and* $\mathbf{b}$ *with minimal approximation effort* $\|[A - \tilde{A}; \mathbf{b} - \tilde{\mathbf{b}}]\|_F$ *such that* $\tilde{\mathbf{b}} \in R(\tilde{A})$ *and* $[A - \tilde{A}; \mathbf{b} - \tilde{\mathbf{b}}]v_3 = 0$.

reason for nonexistence of a TLS solution. Moreover the LS solution exists. Indeed from

$$A = U'\Sigma'V'^T = \begin{bmatrix} \frac{2\sqrt{3}}{\sqrt{13}} & 0 \\ 0 & 1 \\ \frac{1}{\sqrt{13}} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{13} & 0 \\ 0 & .5 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \tag{17}$$

we obtain the LS solution $\mathbf{x}'$ as

$$\begin{aligned} \mathbf{x}' &= V'\Sigma'^{-1}U'^T\mathbf{b} \\ &= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{13}} & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 2\frac{\sqrt{3}}{\sqrt{13}} & 0 & \frac{1}{\sqrt{13}} \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ -\sqrt{3} \end{bmatrix} \\ &= \begin{bmatrix} \frac{3}{13} & \frac{\sqrt{3}}{\sqrt{2}} \\ \frac{3}{13} & \frac{\sqrt{3}}{\sqrt{2}} \end{bmatrix}. \end{aligned} \tag{18}$$

A generalized TLS solution can be obtained by dropping one more singular triplet in (15), i.e., by replacing σ_2 and σ_3 by zero. Then we obtain the TLS approximation $[\hat{A}; \hat{\mathbf{b}}]$ such that $\hat{\mathbf{b}} \in R(\hat{A})$ $\hat{\mathbf{a}}_1 = \hat{\mathbf{a}}_2 = 2\sqrt{3}/2\mathbf{u}_1$, and $\hat{\mathbf{b}} = 2\mathbf{u}_1$. Note that $\hat{\mathbf{b}}$ remains orthogonal to $\mathbf{u}_3$. Now $\hat{\mathbf{b}} \in R(\hat{A})$ and the generalized TLS solution is obtained from (10):

$$\hat{\mathbf{x}}^T = \left[\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right]. \tag{19}$$

As $v_{n+1,n+1} = 0$, $[\hat{\mathbf{x}}^T; -1]^T$ is orthogonal onto $\mathbf{v}_3$, and hence, $[\hat{\mathbf{x}}^T; -1]^T$ is parallel with $\mathbf{v}_2$. We only have to approximate $[A; \mathbf{b}]$, by making only σ_2 in (15) zero, in order to satisfy (3). As the total approximation effort must be minimal, i.e., (4) must be satisfied, σ_3 need not be zeroed. Hence, the TLS approximation which satisfies (3) and (4) such that $[\hat{\mathbf{x}}^T; -1]^T \perp v_3$, is

$$[\tilde{A}; \tilde{\mathbf{b}}] = U \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} V^T \tag{20}$$

and the corresponding correction matrix of rank 1 is given by

$$[\Delta\tilde{A}; \Delta\tilde{\mathbf{b}}] = \sigma_2\mathbf{u}_2\mathbf{v}_2^T = [A; b] \begin{bmatrix} \hat{\mathbf{x}} \\ -1 \end{bmatrix} [\hat{\mathbf{x}}^T; -1] v_{3,2}^2, \tag{21}$$

Observe that $\sigma'_2 = \sigma_3 = .5$ and $\mathbf{u}'_2 = \mathbf{u}_3$. Hence $\mathbf{b} \perp \mathbf{u}'_2$ and $\mathbf{b}'$ remains orthogonal on $\mathbf{u}'_2$. All those facts are generalized and proven in Theorem 3.1. Observe from Fig. 1 that the case where the generic TLS solution does not exist happens only when the length $\|\mathbf{b} - \mathbf{b}'\|_2$ of the orthogonal projection of $\mathbf{b}$ onto $R(A)$ is *larger* than the length σ'_n of the

projection of all columns $\mathbf{a}_i$ of A onto its lowest singular vector. This is the case when A is (nearly) *rank-deficient*, or when the set of equations is *highly incompatible* as in this example.

THEOREM 3.1. *Let* (5) (*respectively*, (6)) *be the* SVD *of* A (*respectively*, $[A; \mathbf{b}]$). *Let* $\mathbf{b}'$ *be the orthogonal projection of* $\mathbf{b}$ *onto* $R(A)$ *and* $[\hat{A}; \hat{\mathbf{b}}]$ *the rank-n approximation of* $[A; \mathbf{b}]$, *as given by* (13). *If* $V'(\sigma_j)$ (*respectively*, $U'(\sigma_j)$) *is the right* (*respectively*, *left*) *singular subspace of* A *associated with* σ_j, *then the following relations can be proven*:

$$
\begin{aligned}
&\text{(a)} && v_{n+1,j} = 0 \Leftrightarrow \mathbf{v}_j = \begin{bmatrix} \mathbf{v}' \\ 0 \end{bmatrix} && \text{with } \mathbf{v}' \in V'(\sigma_j), \\
&\text{(b)} && \Rightarrow \sigma_j = \sigma'_k && \text{with } k = j-1 \quad \text{or} \quad k = j \text{ and } 1 \leqq k \leqq n, \\
&\text{(c)} && \Rightarrow \mathbf{b} \perp \mathbf{u}' && \text{with } \mathbf{u}' \in U'(\sigma_j), \\
&\text{(d)} && \Leftrightarrow \mathbf{u}_j = \mathbf{u}' && \text{with } \mathbf{u}' \in U'(\sigma_j), \\
&\text{(e)} && \Rightarrow \mathbf{b}' \perp \mathbf{u}' && \text{with } \mathbf{u}' \in U'(\sigma_j), \\
&\text{(f)} && \Rightarrow \hat{\mathbf{b}} \perp \mathbf{u}' && \text{with } \mathbf{u}' \in U'(\sigma_j).
\end{aligned} \tag{22}
$$

If σ_j *is an isolated singular value, the converses of relations* (c) *and* (e) *also hold.*

Proof. See [8, p. 36] for the proof. □

As was said before, the generic TLS solution does not exist if $v_{n+1,n+1} = 0$. Theorem 3.1 describes the properties of this situation, namely: if $v_{n+1,n+1} = 0$, then $\sigma_{n+1} = \sigma'_n$. This result can also be derived from Theorem 2.1 for $d = 1$ by contraposition. If additionally $\sigma_n > \sigma_{n+1}$ then

$$
\mathbf{u}_{n+1} = \mathbf{u}'_n, \qquad \mathbf{v}_{n+1} = \begin{bmatrix} \pm \mathbf{v}'_n \\ 0 \end{bmatrix}
$$

and $\mathbf{b}$, $\mathbf{b}'$ as well as $\hat{\mathbf{b}}$ are orthogonal onto $\mathbf{u}'_n$. *For those cases we want to prove now that a* TLS *solution* $\hat{\mathbf{x}}$ *satisfying* (3) *and* (4) *still exists under additional constraints and how it is deduced.* This is done in the next theorem which includes all cases in which the generic TLS solution fails to exist.

THEOREM 3.2. *Let* (6) *be the* SVD *of* $[A; \mathbf{b}]$ *and* $p \leqq n$. *Assume* $v_{n+1,j} = 0$ *for* $j = p+1, \cdots, n+1$ *and* $v_{n+1,p} \neq 0$. *Then, if* $\sigma_{p-1} > \sigma_p$,

$$
[\hat{\mathbf{x}}^T; -1]^T = -\mathbf{v}_p / v_{n+1,p} \tag{23}
$$

is the unique nongeneric TLS *solution satisfying* (4) *subject to* (3) *and* $[\Delta\hat{A}; \Delta\tilde{\mathbf{b}}]v_j = 0$, *for all* j *with*

$$
[\Delta\hat{A}; \Delta\hat{\mathbf{b}}] = \sigma_p \mathbf{u}_p \mathbf{v}_p^T, \tag{24}
$$

$$
[\hat{A}; \hat{\mathbf{b}}] = U\hat{\Sigma}V^T \quad \text{with } \hat{\Sigma} = \operatorname{diag}(\sigma_1, \cdots, \sigma_{p-1}, 0, \sigma_{p+1}, \cdots, \sigma_{n+1}). \tag{25}
$$

Proof. See [11, p. 9] for the proof. □

The *condition* $\sigma_{p-1} > \sigma_p$ *is not a restriction*, and is used here only for reasons of simplicity. Indeed, if $\sigma_{p-1} = \sigma_p$, then the nongeneric TLS solution still exists. We must take the *minimum norm* solution $[\hat{\mathbf{x}}^T; -1]^T$ in the r-dimensional right singular subspace $V(\sigma_p)$ of $[A; \mathbf{b}]$, associated with σ_p of multiplicity r. However, if several $v_{n+1,j} \neq 0$ for $j = p-r+1, \cdots, p$, *the nongeneric* TLS *solution is no longer unique.*

If $v_{n+1,n+1} = 0$, $v_{n+1,n} \neq 0$ and $\sigma_{n-1} > \sigma_n$, the nongeneric TLS solution follows directly from (23) in Theorem 3.2:

$$[\hat{\mathbf{x}}^T; -1]^T = -\mathbf{v}_n / v_{n+1,n} \tag{26}$$

with corresponding correction matrix $[\Delta\hat{A}; \Delta\hat{\mathbf{b}}] = \sigma_n \mathbf{u}_n \mathbf{v}_n^T$ having minimal Frobenius norm σ_n such that $(\mathbf{b} - \Delta\hat{\mathbf{b}}) \in R(A - \Delta\hat{A})$ and $[\Delta\hat{A}; \Delta\hat{\mathbf{b}}]v_{n+1} = 0$.

Let us also *compare* LS *and* TLS *in the nongeneric case.* LS looks for a solution $\mathbf{x}'$ such that $\sum_{i=1}^{n} \sigma_i' \mathbf{u}_i' \mathbf{v}_i'^T \mathbf{x}' = \mathbf{b}'$. If

$$\mathbf{v}_j = \begin{bmatrix} \mathbf{v}_j'' \\ v_{n+1,j} \end{bmatrix}$$

with $v_{n+1,j} = 0$ for $j = p+1, \cdots, n+1$, then by Theorem 3.1, this implies that $\sigma_{j-1}' = \sigma_j$ and $\mathbf{b}$, as well as $\mathbf{b}'$, are orthogonal onto $\mathbf{u}_j = \mathbf{u}'$ with $\mathbf{u}' \in U'(\sigma_j)$ for $j = p+1, \cdots, n+1$. Or, $[A; \mathbf{b}]_p = [A_{p-1}; \mathbf{b}]$ with $[A; \mathbf{b}]_p$ (respectively, A_{p-1}) the rank-p (respectively, rank-$(p-1)$) approximation of $[A; \mathbf{b}]$ (respectively, A), as derived from the Eckart–Young theorem [6, p. 19]. Also, $\mathbf{x}' \perp \mathbf{v}_j''$, $j = p+1, \cdots, n+1$. Hence, LS must also look for a solution $\mathbf{x}'$ in the $(p-1)$-dimensional row space $R_R(A_{p-1})$ of A_{p-1}, orthogonal onto span $\{\mathbf{v}_j'' \mid j = p+1, \cdots, n+1\}$, which equals span $\{\mathbf{v}_j' \mid j = p, \cdots, n\}$. Thus $\mathbf{b}$, as well as $\hat{\mathbf{b}}$ and $\mathbf{b}'$, is orthogonal onto the same $n-p+1$ singular vectors $\mathbf{u}_j$, $j = p+1, \cdots, n+1$. Both TLS and LS search for an optimal solution in a subspace of same dimension $p-1$ with p the largest index such that $v_{n+1,p} \neq 0$, i.e., dim $(R[A_{p-1}; \mathbf{b}']) =$ dim $(R[A_{p-1}; \mathbf{b}]_{p-1}) = p-1$. Hence, it follows that

$$\begin{aligned} \|\mathbf{b} - \mathbf{b}'\|_2 = \|[A_{p-1}; \mathbf{b}] - [A_{p-1}; \mathbf{b}']\|_F &\geqq \min_{\text{rank}(C) = p-1} \|[A_{p-1}; \mathbf{b}] - C\|_F \\ &= \sigma_p = \|[\Delta\hat{A}; \Delta\hat{\mathbf{b}}]\|_F \end{aligned} \tag{27}$$

with $[\Delta\hat{A}; \Delta\hat{\mathbf{b}}]$ given by (24). The LS *approximation effort remains larger* than the TLS approximation effort. Observe that the *generic* TLS solution *does not exist*, if the set of equations $A\mathbf{x} \approx \mathbf{b}$ is *highly incompatible* (so that $\|\mathbf{b} - \mathbf{b}'\|_2$ is large and exceeds σ_p' of A), or if A is *nearly rank-deficient* (i.e., $\sigma_p' \approx \cdots \approx \sigma_n' \approx 0$).

In the extreme case, where $\mathbf{b}$ is orthogonal on $R(A)$, TLS and LS give the following zero solution: as $\mathbf{b} \perp R(A)$, $\mathbf{b}' = 0$. Hence, $\mathbf{x}'$ is zero. The zero nongeneric TLS solution follows from direct application of Theorem 3.2. Indeed, the set of left singular vectors of $[A; \mathbf{b}]$ is precisely $\{\mathbf{u}_1', \mathbf{u}_2', \cdots, \mathbf{b}/\|\mathbf{b}\|_2, \mathbf{u}_i', \cdots, \mathbf{u}_n'\}$. This is the set of left singular vectors of A, where $\mathbf{b}$ is introduced at the ith location corresponding to its size $\sigma_{i-1}' \geqq \|\mathbf{b}\|_2 \geqq \sigma_i'$. Correspondingly, the ith column $\mathbf{v}_i$ of V is $[0, \cdots, 0, 1]$ and the last row of V is $[0, \cdots, 0, 1, \cdots, 0]$ with its ith element equal to one. Since $v_{n+1,j} = 0$ for $j = i+1, \cdots, n+1$ and $v_{n+1,i} \neq 0$, the TLS solution $[\hat{\mathbf{x}}^T; -1]^T$ is given by $-\mathbf{v}_i / v_{n+1,i}$. Hence, $\hat{\mathbf{x}}$ is zero.

In practice $v_{n+1,j}$ will seldom be zero. Indeed, due to errors in the observations $[A; \mathbf{b}]$, a zero-valued $v_{n+1,j}$ in the SVD of the unobservable exact data matrix $[A_0; \mathbf{b}_0]$, will differ from zero in the SVD of $[A; \mathbf{b}]$. Hence it is advisable to define an error bound ε such that all $|v_{n+1,j}| < \varepsilon$ are considered to be zero.

3.2. The multidimensional case. Before describing the properties of this situation, we first *prove that a* TLS *solution* $\hat{X}$ satisfying (3) and (4) *still exists* under additional

constraints and how it is deduced. This is done in the next theorem, which generalizes Theorem 3.2 and *includes all cases* in which the *generic* multidimensional TLS *solution fails to exist.*

THEOREM 3.3. *Let* (6) *be the* SVD *of* $[A; B]$. *Assume* $\sigma_p > \sigma_{p+1} = \cdots = \sigma_{n+1}$. *Let* Q_1 *be an orthonormal matrix such that*

$$[\mathbf{v}_{p+1}, \cdots, \mathbf{v}_{n+d}]Q_1 = \begin{bmatrix} Y_1 & Z_1 \\ 0 & \Gamma_1 \end{bmatrix} \begin{matrix} n \\ d \end{matrix} \qquad \begin{matrix} n-p & d \end{matrix} \tag{28}$$

with Γ_1 *singular of corank* κ *and* $p \leqq n$. *Let* Q *be an orthonormal matrix such that*

$$[\mathbf{v}_{q+1}, \cdots, \mathbf{v}_{n+d}]Q = \begin{bmatrix} Y & Z \\ 0 & \Gamma \end{bmatrix} \begin{matrix} n \\ d \end{matrix} \qquad \begin{matrix} n-q & d \end{matrix} \tag{29}$$

with minimal $n - q$ *such that* Γ *is nonsingular* ($p - q \geqq \kappa$). *Assume* $\sigma_q > \sigma_{q+1}$. *Then*

$$\hat{X} = -Z\Gamma^{-1} \tag{30}$$

is the unique nongeneric TLS *solution satisfying* (4) *subject to* (3) *and* $[\Delta\hat{A}; \Delta\hat{B}][\begin{smallmatrix}\mathbf{w}\\0\end{smallmatrix}]_d^n = 0$ *for all* $[\begin{smallmatrix}\mathbf{w}\\0\end{smallmatrix}] \in \operatorname{span}\{\mathbf{v}_{q+1}, \cdots, \mathbf{v}_{n+d}\}$

$$[\Delta\hat{A}; \Delta\hat{B}] = [A; B]\begin{bmatrix} Z \\ \Gamma \end{bmatrix}[Z^T; \Gamma^T]. \tag{31}$$

Proof. See [11, p. 11] for the proof. □

Theorem 3.3 *proves* that *a* TLS *solution*, called *nongeneric*, satisfying the TLS criteria (3) and (4), still *exists* under additional constraints. Theorem 3.3 also proves how it is computed: if the corank κ of V_{22} or Γ_1 is $\geqq 1$, we must add the minimum number (at least κ) of right singular vectors associated with the smallest singular values of $[A; B]$ until a nonsingular Γ in (29) is obtained. The basis $[\begin{smallmatrix}Z\\\Gamma\end{smallmatrix}]$ is orthogonal onto the remaining independent base vectors $[\begin{smallmatrix}Y\\0\end{smallmatrix}]$ of span $\{\mathbf{v}_{q+1}, \cdots, \mathbf{v}_{n+d}\}$ which make Γ singular.

As in the one-dimensional case (see Theorem 3.2), the *condition* $\sigma_q > \sigma_{q+1}$ *is not a restriction* and is only used here for reasons of simplicity. Indeed, if $\sigma_q = \sigma_{q+1}$ then the nongeneric TLS solution still exists and we must take the *minimum norm* solution.

In the following theorems, we describe the *properties in the case of singularity* of V_{22} (or Γ_1) by *generalizing* the properties given in Theorem 3.1 to the multidimensional case. Observe that *additional conditions* concerning the multiplicity of σ_{n+1} are imposed.

Let us first assume that the *unique generic multidimensional* TLS *solution does not exist*, i.e., $\sigma_n > \sigma_{n+1}$ and V_{22} is singular.

THEOREM 3.4. *Let* (5) (*respectively*, (6)) *be the* SVD *of* A (*respectively*, $[A; B]$). *Let* $\sigma_n > \sigma_{n+1}$ *and* V_{22} *be singular. Call* $V'(\sigma_j)$ (*respectively*, $U'(\sigma_j)$) *the right* (*respectively*, *left*) *singular subspace of* A *associated with* σ_j, *and* $V(\sigma_j)$ (*respectively*, $U(\sigma_j)$) *the right* (*respectively*, *left*) *singular subspace of* $[A; B]$, *associated with* σ_j. *Consider any unit vector* $\mathbf{x} \in \operatorname{Ker}(V_{22})$ *and let* b (*respectively*, r) *be the index of the first* (*respectively*, *last*) *nonzero component of* $\mathbf{x}$. *Then the following relations can be proven. If* $\sigma_{n+b} = \cdots = \sigma_j = \cdots = \sigma_{n+r}$, *then*

$$
\begin{aligned}
&\text{(a)} \quad \exists\, \mathbf{v} \in V(\sigma_j), \quad \exists\, \mathbf{v}' \in V'(\sigma_j): \begin{bmatrix} V_{12} \\ V_{22} \end{bmatrix} \mathbf{x} = \mathbf{v} = \begin{bmatrix} \mathbf{v}' \\ 0 \end{bmatrix}, \\
&\text{(b)} \quad \exists\, i: \sigma_j = \sigma_i', \\
&\text{(c)} \quad \exists\, \mathbf{u}' \in U'(\sigma_j): B \perp \mathbf{u}', \\
&\text{(d)} \quad \exists\, \mathbf{u} \in U(\sigma_j), \quad \exists\, \mathbf{u}' \in U'(\sigma_j): \mathbf{u} = \mathbf{u}', \\
&\text{(e)} \quad \exists\, \mathbf{u}' \in U'(\sigma_j): B' \perp \mathbf{u}', \\
&\text{(f)} \quad \exists\, \mathbf{u}' \in U'(\sigma_j): \hat{B} \perp \mathbf{u}'.
\end{aligned} \tag{32}
$$

Proof. See [8, p. 47] for the proof. □

If all components of $\mathbf{x} \in \text{Ker}\,(V_{22})$ differ from zero, the additional conditions are $\sigma_{n+1} = \cdots = \sigma_{n+d}$. In this case, every linear combination of the columns of $[{}^{V_{12}}_{V_{22}}]$ is a right singular vector of $[A; B]$, associated with σ_{n+1}.

For the one-dimensional case ($d = 1$) with $\sigma_n > \sigma_{n+1}$ and $v_{n+1,n+1} = 0$, no additional conditions are required and Theorem 3.4 reduces to a special case of Theorem 3.1.

Assuming that the conditions of Theorem 3.4 are satisfied, let us consider the *properties of the nongeneric* TLS *solution* (30), as deduced in Theorem 3.3. Assume that a nonsingular Γ can be obtained in (29) of Theorem 3.3 for $q = n - 1$ and $\sigma_{n-1} > \sigma_n$. In this case, corank $(V_{22}) = 1$. Consider $\sigma'_{n-1} > \sigma'_n$; then Theorem 3.4 yields

$$
\begin{aligned}
&\exists\, \mathbf{v} \in V(\sigma_j), \quad \exists\, \mathbf{u} \in U(\sigma_j): \begin{bmatrix} V_{12} \\ V_{22} \end{bmatrix} \mathbf{x} = \mathbf{v} = \begin{bmatrix} \mathbf{v}'_n \\ 0 \end{bmatrix}, \quad \sigma_j = \sigma'_n, \quad \mathbf{u} = \mathbf{u}'_n, \\
&B (\text{and } \hat{B}) \perp \mathbf{u} = \mathbf{u}'_n \in U(\sigma_j) \quad \text{and} \quad n+1 \leqq j \leqq n+d.
\end{aligned} \tag{33}
$$

Hence, B cannot be approximated in the space generated by the left singular vectors $\{\mathbf{u}_{n+1}, \cdots, \mathbf{u}_{n+d}\}$ of $[A; B]$ such that (3) is satisfied. This means that there is no generic TLS solution along $\mathbf{v}$ (see (33)). Hence, we then look for a nongeneric $[\hat{X}^T; -I_d]^T$ which is orthogonal onto $\mathbf{v}$. The correction matrix, given by (31) and satisfying $[\Delta\hat{A}; \Delta\hat{B}]\mathbf{v} = 0$, has minimal norm given by

$$
\|[\Delta\hat{A}; \Delta\hat{B}]\|_F = \sqrt{\sum_{\substack{k=n \\ k \neq j}}^{n+d} \sigma_k^2}. \tag{34}
$$

As in the one-dimensional case, we can also compare the nongeneric TLS solution with the LS solution. The same conclusions hold for multidimensional problems (1) (see [8, p. 50]).

If corank $(V_{22}) = \kappa \geqq 1$, Theorem 3.4 can be applied to each vector $\in \text{Ker}\,(V_{22})$. Or, more conveniently, Theorem 3.4 can be formulated for κ-dimensional subspaces (see [8, p. 51]).

Also, Theorem 3.4 can be straightforwardly extended to describe the properties of the case that the minimum norm generic TLS solution (8) cannot be computed with the algorithm of Golub and Van Loan [5], i.e., $\sigma_{n-p} > \sigma_{n-p+1} = \cdots = \sigma_{n+1}$, $p \geqq 1$ and Γ in (9) singular (see [8, p. 53]).

The *additional conditions* may appear somehow restrictive. However, *solving* the set of equations $AX \approx B$ *with* TLS *makes sense only when the data A and B are observations of an exact but unobservable relation* $A_0X = B_0$. Due to errors, the singular values $\sigma_{n+1}, \cdots, \sigma_{n+d}$ of $[A; B]$ will differ from the exact singular values $\sigma^0_{n+1} = \cdots = \sigma^0_{n+d} = 0$ of $[A_0; B_0]$. Hence, it is advisable to consider $\sigma_{n+1}, \cdots, \sigma_{n+d}$ of $[A; B]$ as *coinciding*, and thus assume that the additional conditions of Theorem 3.4 are indeed satisfied.

If V_{22} (or Γ) is singular and the *additional conditions are not satisfied*, the nongeneric TLS solution still exists as proven in Theorem 3.3. However, for those cases the nice properties described in Theorem 3.4 are no longer valid and no adequate comparison with the LS solution can be made. This situation will occur if at least one subset $A\mathbf{x}_i \approx \mathbf{b}_i$ in (1) is highly incompatible. It affects the accuracy of the other solutions when solving the d-dimensional TLS problem. In this case, it is expected to obtain more accurate TLS solutions $\hat{\mathbf{x}}_i$, $i = 1, \cdots, d$, by solving d single one-dimensional TLS problems $A\mathbf{x}_i \approx \mathbf{b}_i$, $i = 1, \cdots, d$.

4. Outline of the generalized TLS algorithm. In this section, we want to summarize the practical TLS computation into a *generalized* algorithm which handles the generic TLS problem of any dimension d, as well as the case of nonuniqueness and the nongeneric case.

ALGORITHM 4.1. Computation of the TLS solution $\hat{X}$ of $AX \approx B$.
Given: an m by n data matrix A and an m by d observation matrix B.
Step 1.
1(a). If $m > 5/3(n + d)$, transform $[A; B]$ into upper triangular form R by Householder transformations.
1(b). Compute the singular value decomposition (6) of $[A; B]$ (or R).
Step 2. Compute the rank r ($\leqq n$) of $[A; B]$ by

(35) $\sigma_1^2 \geqq \cdots \geqq \sigma_r^2 > R_0 \geqq \sigma_{r+1}^2 \geqq \cdots \geqq \sigma_{n+d}^2$ with R_0 an appropriate rank determinator.

Step 3. Compute by Householder transformations an orthogonal matrix Q such that

(36) $$[\mathbf{v}_{r+1}, \cdots, \mathbf{v}_{n+d}]Q = \begin{bmatrix} & & Z \\ & & \\ 0 & & \Gamma \end{bmatrix} \begin{matrix} n \\ \\ d \end{matrix}$$
$$\qquad\qquad\qquad\qquad\qquad\quad n-r \qquad d$$

with Γ a d by d upper triangular matrix.

Step 4. If Γ is singular then begin:

lower the rank r with the multiplicity of σ_r
go back to Step 3
end
(37) else solve by forward elimination: $\hat{X}\Gamma = -Z$

END

A fully documented Fortran program of this TLS algorithm is given in [7].

If $m < n$, the set of equations $AX = B$ is always *underdetermined.* This implies that the solution, generically, is never unique. In this case, Algorithm 4.1 can still be applied to compute the minimum norm solution.

If $d = 1$, the rank $r = n$, $|\Gamma| > \varepsilon > 0$ and $\sigma_n > \sigma_{n+1}$, (37) reduces to (12). The TLS solution is obtained from a simple scaling of the right singular vector $\mathbf{v}_{n+1}$ of $[A; \mathbf{b}]$, associated with its minimal singular value σ_{n+1}.

If $m > 5/3(n + d)$, it is more efficient to first *triangularize* $[A; B]$, using a QR factorization [2], and then proceed to R. Indeed, $[A; B]$ and R have the same singular values and right singular vectors. Hence, the TLS solution $\hat{X}$ of $AX \approx B$ can also be obtained from the SVD of R. This option has been incorporated into Algorithm 4.1 (see [7]).

The upper triangular matrix Γ must be *tested for nonsingularity.* This is done by comparing the absolute value of each diagonal element $|\Gamma_{ii}|$, $i = 1, \cdots, d$ with a given small positive number ε. The choice of ε can be based on the numerical accuracy of the SVD computation [4] or, preferably, can depend upon the perturbations of the data [12]. A generally applicable, reliable error bound ε in function of the standard deviation σ_v of the perturbations has not yet been deduced.

Observe from (36) that we *need only to compute a few singular vectors* associated with the smallest singular values of $[A; B]$ in order to obtain the TLS solution $\hat{X}$. Moreover, we *need only to compute a basis of the singular subspace corresponding to the smallest* $n + d - r$ *singular values with* r *the rank of* $[A; B]$. Indeed, the TLS solution $\hat{X}$ is *invariant* with respect to orthogonal base transformations in its solution space $\mathrm{R}([\begin{smallmatrix}\hat{X}\\ -I_d\end{smallmatrix}])$ (see [8, p. 57]).

On the basis of these properties we were able to *improve the efficiency* of the TLS computations by *computing the* SVD of $[A; B]$ in Step 1 only "partially." This results in the development of a *new algorithm* "partial total least squares (PTLS)." PTLS is about two times faster than the classical TLS computation given in Algorithm 4.1, while the same accuracy can be maintained. For more details, we refer the reader to [8]–[10].

5. Conclusion. In this paper, the total least squares problem, and the classical algorithm of Golub and Van Loan used to solve it, are generalized to all nongeneric cases, i.e., problems in which *the algorithm of Golub and Van Loan fails to produce a* TLS *solution* (§ 2). We prove that under additional constraints, the proposed generalization remains *optimal* with respect to the TLS criteria in the one-dimensional (§ 3.1) as well as in the multidimensional case (§ 3.2) and describe the properties of those problems. It is concluded that nongeneric TLS problems occur when the set of equations $AX \approx B$ is highly incompatible or when the data matrix A is (nearly) rank-deficient.

Finally, the TLS computations are summarized in one algorithm which includes the proposed generalization (§ 4).

REFERENCES

[1] R. J. ADCOCK, *A problem in least squares,* The Analyst, 5 (1878), pp. 53–54.

[2] T. F. CHAN, *An improved algorithm for computing the singular value decomposition,* ACM Trans. Math. Software, 8 (1982), pp. 72–83.

[3] L. J. GLESER, *Estimation in a multivariate "errors in variables" regression model: Large sample results,* Ann. Statist., 9 (1981), pp. 24–44.

[4] G. H. GOLUB AND C. REINSCH, *Singular value decomposition and least squares solutions,* Numer. Math., 14 (1970), pp. 403–420.

[5] G. H. GOLUB AND C. F. VAN LOAN, *An analysis of the total least squares problem,* SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[6] ———, *Matrix computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.
[7] S. VAN HUFFEL, *Documented Fortran 77 programs of the extended classical total least squares algorithm, the partial singular value decomposition algorithm and the partial total least squares algorithm*, Internal report ESAT-KUL 88/1, Dept. of Electrical Engineering, Katholieke Univ. Leuven, February 1988.
[8] ———, *Analysis of the total least squares problem and its use in parameter estimation*, Ph.D. thesis, Dept. of Electrical Engineering, Katholieke Univ. Leuven, June 1987.
[9] S. VAN HUFFEL, J. VANDEWALLE, AND A. HAEGEMANS, *An efficient and reliable algorithm for computing the singular subspace of a matrix, associated with its smallest singular values*, J. Comput. Appl. Math., 19 (1987), pp. 313–333.
[10] S. VAN HUFFEL AND J. VANDEWALLE, *The partial total least squares algorithm*, J. Comput. Appl. Math., 21 (1988), pp. 333–341.
[11] ———, *Comments on the solution of the nongeneric total least squares problem*, Internal report ESAT-KUL 88/3, Dept. of Electrical Engineering, Katholieke Univ. Leuven, March 1988.
[12] P. A. WEDIN, *Perturbation bounds in connection with the singular value decomposition*, BIT, 16 (1972), pp. 99–111.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

008

SELF-DUAL POLYNOMIALS, HERMITE MATRICES, AND HEISENBERG FUNCTIONALS*

RON PERLINE†

Abstract. For certain orthogonal matrices associated with classical self-dual discrete orthogonal families, commuting symmetric tridiagonal matrices are constructed. Their eigenvectors are shown to be critical points for functionals related to Heisenberg's inequality.

Key words. orthogonal polynomials, commuting tridiagonal matrices, discrete Fourier transform

AMS(MOS) subject classification. 33A70

1. Introduction. We begin by recalling three fundamental facts from elementary Fourier analysis [DM]:

(i) Heisenberg's inequality. For $f \in L^2(R^1)$, we have

$$\left(\int x^2 |f(x)|^2 \, dx\right)\left(\int x^2 |Ff(x)|^2 \, dx\right) \geqq \frac{1}{4} \|f\|^4,$$

where $\|f\|$ is the L^2 norm of f and

$$Ff(x) = (2\pi)^{-1/2} \int f(t) e^{-ixt} \, dt$$

is the Fourier transform of f. The lower bound is attained only if f is Gaussian, that is, $f(x) = \exp(-kx^2)$, $k > 0$. Equality holds in particular for the Gaussian $\psi(x) = \exp(-x^2/2)$.

(ii) Eigenvector decomposition for the Hermite operator. The Gaussian ψ is the "ground state" (eigenvector corresponding to the lowest eigenvalue) of the Hermite operator $L\psi = -\psi'' + x^2\psi$. In general, the eigenvectors for L are of the form $H_n(x) \exp(-x^2/2)$, where $H_n(x)$ is the nth Hermite polynomial.

(iii) Commutativity. The Hermite operator L commutes with the Fourier transform F. L has simple spectrum; thus the eigenvectors of L are eigenvectors of F.

The theme of this paper is that these facts are tightly intertwined. We will consider discrete analogues of the Fourier transform associated with self-dual discrete orthogonal polynomial families. For each such Fourier transform analogue F we construct a self-adjoint, second-order difference operator L which *commutes* with F. Thus the eigenvectors of L are again shared by F. These common eigenvectors enjoy another property: they are critical points for a functional closely related to the Heisenberg inequality. Because of the analogy with the case of ordinary Fourier analysis, we refer to L as a *Hermite* matrix.

The construction of L is inspired by previous work of Grunbaum [Gr] and Dickinson and Steiglitz [DS]. In both cases, difference operators commuting with the discrete Fourier transform are obtained. One motivation for studying these commuting difference operators is that their eigenvectors form a canonical basis with respect to which the discrete Fourier transform is diagonal. Thus it is plausible that there may exist fast transform algorithms, which would utilize the existence of these eigenvectors in some way. To date, this hope has not been realized.

* Received by the editors May 8, 1987; accepted for publication October 23, 1987.

† Department of Mathematics and Computer Science, Drexel University, Philadelphia, Pennsylvania 19104.

We remark that the commuting operators of [Gr] and [DS] differ in that they correspond to discrete Dirichlet and periodic boundary conditions, respectively.

2. Classical self-dual families. We will be considering the following discrete orthogonal families (see [A]):

(1) Poisson–Charlier polynomials:

$$c_n(x) = {}_2F_0(-n, -x; -; -1/a), \quad a > 0, \quad x = 0, 1, 2, \cdots,$$

$$w(x) = e^{-a}a^x/x!$$

(2) Meixner polynomials:

$$M_n(x) = {}_2F_1(-n, -x; \beta; 1 - 1/c), \quad \beta > 0, \quad 0 < c < 1,$$

$$w(x) = c^x(\beta)_x/x!, \qquad x = 0, 1, 2, \cdots.$$

(3) Krawtchouk polynomials:

$$K_n(x) = {}_2F_1(-n, -x; -N; 1/p), \qquad 0 < p < 1,$$

$$w(x) = \binom{N}{x} p^x(1-p)^{N-x}, \qquad x = 0, 1, 2, \cdots, N.$$

In each case, we have recorded the weight with respect to which the polynomials are orthogonal.

These families are self-dual, which for our purposes simply means that $p_i(j) = p_j(i)$ for any polynomial in the families listed. The fact that our polynomials enjoy this property is immediate from their defining formulae.

As defined above, the polynomials are orthogonal but not normal. The relevant inner product formulae are given in [A, p. 15]. In each case, the formula can be written:

$$\sum (p_i(x))^2 w(x) = d/w(i) \quad \text{where } d \text{ is a positive value,}$$

depending perhaps upon auxiliary quantities, but *independent* of i. This will be important later on. We define c, the normalization constant, to be equal to $\sqrt{d}$.

It is well known that these polynomials are in each case eigenvectors of second-order difference operators, self-adjoint with respect to the weights $w(x)$. Let $\Delta_+ f(x) = f(x+1) - f(x)$, $\Delta_- f(x) = f(x) - f(x-1)$. We list the associated difference operators [L], [P]:

(1) Poisson–Charlier:

$$e^a a^{-x} x! \Delta_+ [e^{-a} a^{x-1}/(x-1)!(-a)\Delta_- c_i(x)] = i c_i(x).$$

(2) Meixner:

$$c^{-x} x!/(\beta)_x \Delta_+ [c^{x-1}(\beta)_x/(x-1)!(x-1+\beta)\Delta_- M_i(x)] = i(1-1/c)M_i(x).$$

(3) Krawtchouk:

$$1/w(x)\Delta_+ [(x-1-N)pw(x-1)\Delta_- K_i(x)] = iK_i(x).$$

3. Matrix reformulation. We restate the properties described above in matrix notation. Let $\{p_i\}$ be any one of our orthogonal families, and let

$$F_1 = (p_i(j));$$

that is, F_1 is the matrix whose (i, j)th entry is $p_i(j)$. F_1 is semi-infinite in the case of Poisson–Charlier and Meixner. Observe that F_1 is in any case a symmetric matrix. Let

$$\Lambda = (\Lambda_{ij}), \qquad \Lambda_{ij} = \sqrt{w(i)}\,\delta_{ij}.$$

Then the orthogonality relations can be restated as

(A) $F_1\Lambda^2F_1^t = F_1\Lambda^2F_1 = c^{-2}\Lambda^{-2}$.

The fact that the polynomials are eigenvectors of a second-order difference operator can be written as

(B) $\Lambda_2F_1 = F_1D_1$

where D_1 is a tridiagonal matrix and Λ_2 is the diagonal matrix whose entries are the eigenvalues of the difference operator. The self-adjointness of the difference operator with respect to the weight w simply corresponds to the fact that D_1 satisfies

(C) $D_1\Lambda^2 = \Lambda^2D_1^t$.

As an immediate consequence of (A), we have $(\Lambda F_1\Lambda)(\Lambda F_1\Lambda) = c^2I$. Thus if we define $F = c^{-1}(\Lambda F_1\Lambda)$, then F is symmetric *and* orthogonal.

From (B), we derive

$$\Lambda_2\Lambda F_1\Lambda = \Lambda F_1D_1\Lambda = \Lambda F_1\Lambda\Lambda^{-1}D_1\Lambda$$

$$\Rightarrow \Lambda_2F = F\Lambda^{-1}D_1\Lambda.$$

Define $D = \Lambda^{-1}D_1\Lambda$. We can use (C) to obtain $\Lambda^{-1}D_1\Lambda = \Lambda D_1^t\Lambda^{-1}$, which says that D is symmetric. This gives the following implications, by transposition:

$$\Lambda_2F = FD \Rightarrow F^t\Lambda_2 = D^tF^t \Rightarrow DF = F\Lambda_2.$$

Adding the first and third equalities, we finally obtain $(\Lambda_2 + D)F = F(D + \Lambda_2)$. Thus $L = \Lambda_2 + D$ *commutes* with F. We refer to L as the Hermite matrix associated with the discrete orthogonal family.

We give explicit formulae for the matrices associated with the Krawtchouk polynomials. For convenience, matrix entries are parameterized by $0 \leqq i, j \leqq N$. Let $q = 1 - p$. Then we have the following:

(i) $F_1(i, j) = {}_2F_1(-i, -j; -N; 1/p)$.

(ii) $w(i) = \binom{N}{i}p^i(1 - p)^{N-i}$ and Λ is diagonal with $\Lambda_{ii} = \sqrt{w(i)}$.

(iii) c = normalization constant = $q^{N/2}$.

(iv) $F(i, j) = q^{-N/2}\sqrt{w(i)w(j)}F_1(i, j)$.

(v) D_1 is tridiagonal with entries on the ith row

$$(\cdots -(N - i + 1)p\ p(N - 2i) + i\ -(i + 1)q\cdots).$$

(vi) Λ_2 is diagonal with $(\Lambda_2)_{ii} = i$.

(vii) D is tridiagonal with entries on the ith row

$$(\cdots -\sqrt{(N - i + 1)(i)pq}\ p(N - 2i) + i\ -\sqrt{(N - i)(i + 1)pq}\ \cdots).$$

(viii) $L = D + \Lambda_2$.

We now change gears a bit and consider an extremal problem.

4. An extremal problem. Let A and B be two positive, self-adjoint operators on an inner product space V; we wish to find extrema of the functional $(Av, v)(Bv, v)/(v, v)^2$ for $v \neq 0$. Because of the homogeneity of the functional, this is equivalent to the problem of finding extrema of $(Av, v)(Bv, v)$ subject to $(v, v) = c$. We ask under what conditions are such extrema v simultaneously eigenvectors for the operator $L = A + B$?

For example, let $V = L^2(R^1)$, $Af(x) = x^2 f(x)$, $Bf(x) = -f''(x)$. Then it is easily seen that

$$\int x^2 |f(x)|^2 \, dx \int x^2 |Ff(x)|^2 \, dx = (Af, f)(Bf, f).$$

Thus the functional in Heisenberg's inequality is the model for our extremal problem.

To solve our problem, we use Lagrange multipliers. Let $g(v) = (v, v)$, and $h(v) = (Av, v)(Bv, v)$. We wish to find extrema of h subject to the constraint $g(v) = c$. Thus the equations for Lagrange multipliers are (i) grad $h = \lambda$ grad g; (ii) $g = c$. For our particular problem, (i) says $[Av(Bv, v) + Bv(Av, v)] = \lambda v$. Assuming that $(A + B)v = \mu v$, we have $Bv = (-A + \mu I)v$ which implies

$$Av((-A + \mu I)v, v) + (-A + \mu I)v(Av, v) = \lambda v.$$

We collect terms

$$Av\{((-A + \mu I)v, v) - (Av, v)\} = \{\lambda - \mu(Av, v)\}v$$

or

$$(*) \qquad Av\{\mu(v, v) - 2(Av, v)\} = \{\lambda - \mu(Av, v)\}v.$$

Thus, either (i) v is an eigenvector for A, or (ii) $\mu(v, v) - 2(Av, v) = 0$. Let us investigate the consequences of (i).

If v is an eigenvector for A, and also for $A + B$, then v is also an eigenvector for B. In our applications, A and B will have no common eigenvectors.

Supposing (ii) holds, *define* $\lambda = \mu(Av, v)$; then $(*)$ holds and so v is a critical point for our problem.

5. The Heisenberg functional for self-dual polynomials. From our discussion of self-dual polynomials, we have $\Lambda_2 F = FD$, where F, D, Λ_2 are symmetric, F is orthogonal, Λ_2 is diagonal, and D is tridiagonal. We wish to use the computations of the previous section to find critical points for the *Heisenberg functional* $(\Lambda_2 v, v)(Dv, v)/(v, v)^2$. We know that v will simultaneously be a critical point for this functional and satisfy $(\Lambda_2 + D)v = \mu v$ (here we are letting $A = \Lambda_2$ and $B = D$) if μ satisfies $\mu = 2(\Lambda_2 v, v)/(v, v)$. But

$$\Lambda_2 v + Dv = \mu v \Rightarrow (\Lambda_2 v, v) + (Dv, v) = \mu(v, v).$$

Also, since $\Lambda_2 F = FD$, we have $D = F^{-1}\Lambda_2 F = F\Lambda_2 F$. We use this last relation to substitute for D:

$$(\Lambda_2 v, v) + (F\Lambda_2 Fv, v) = \mu(v, v) \Rightarrow (\Lambda_2 v, v) + (\Lambda_2 Fv, Fv) = \mu(v, v).$$

Since F commutes with $L = \Lambda_2 + D$, v is *also* an eigenvector of F. Since $F^2 = I$, the eigenvalues of F are $\tau = \pm 1$, so we have $(\Lambda_2 v, v) + \tau^2(\Lambda_2 v, v) = \mu(v, v) \Rightarrow 2(\Lambda_2 v, v) = \mu(v, v)$, the desired relation.

6. Concluding remarks. For each of the classical families of self-dual discrete polynomials, we have exhibited an associated symmetric orthogonal matrix F which has a commuting symmetric tridiagonal matrix L, the associated Hermite matrix. The eigenvectors of L are critical points of a Heisenberg-type functional associated with each polynomial family.

Similar results hold for the discrete Fourier transform matrix

$$F = 1/\sqrt{N}(\omega^{(i-1)(j-1)}), \qquad \omega^N = 1.$$

The eigenvectors of the operator L of Dickinson and Steiglitz are critical points for the functional $(Av, v)(Bv, v)$, where A is the cyclic matrix

$$A = \begin{pmatrix} -2 & 1 & & 1 \\ 1 & -2 & 1 & \\ & 1 & -2 & 1 \\ 1 & & 1 & -2 \end{pmatrix}$$

and B is the diagonal matrix with entries $\lambda_i = 2(\cos(2\pi(i-1)/N) - 1)$, $i = 1, 2, \cdots, N$. As usual, the Hermite matrix L equals $A + B$.

For the Hermite matrices L associated with our self-dual families, or the Hermite matrix associated with the discrete Fourier transform, it would be of some interest to obtain explicit formulae for the eigenvectors and eigenvalues. The fact that these eigenvectors have some special properties (critical points for some nontrivial functional) encourages us to believe this may be feasible. We are currently investigating this problem.

REFERENCES

[A] R. ASKEY, *Orthogonal Polynomials and Special Functions*, CBMS–NSF Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1975.

[DS] B. DICKINSON AND K. STEIGLITZ, *Eigenvectors and functions of the discrete Fourier transform*, IEEE Trans. Acoust. Speech Signal Process., ASSP-30 (1982), pp. 25–31.

[DM] H. DYM AND H. MCKEAN, *Fourier Series and Integrals*, Academic Press, New York, 1972.

[Gr] F. A. GRUNBAUM, *The eigenvectors of the discrete Fourier transform: a version of the Hermite functions*, J. Math. Anal. Appl., 88 (1982), pp. 355–363.

[L] R. LESKY, *Orthogonale Polynomsysteme als losungen Sturm–Liouvillescher Differenzengleichungen*, Monatsh. Math., 66 (1962), pp. 203–214.

[P] M. PERLSTADT, *Chopped orthogonal polynomial expansions—some discrete cases*, SIAM J. Algebraic Discrete Methods, 4 (1983), pp. 94–100.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

009

DUAL ALGORITHMS FOR ORTHOGONAL PROCRUSTES ROTATIONS*

A. SHAPIRO† AND J. D. BOTHA†

Abstract. This paper considers a problem of rotating m matrices toward a best least-squares fit. The problem is known as the orthogonal Procrustes problem. For $m = 2$ the solution of this problem is known and can be given in a closed form using the singular value decomposition. It appears that the general case of $m > 2$ cannot be solved explicitly and an iterative procedure is required. The authors discuss a dual approach to the Procrustes problem where the maximal value of the objective function is approximated from above. This involves minimization of the sum of k largest eigenvalues of a symmetric matrix. It will be shown that under certain conditions ensuring differentiability of the obtained function at the minimum, this method gives the global solution of the Procrustes problem.

Key words. orthogonal rotation, best least-squares fit, singular value decomposition, least upper bound, eigenvalues, nonsmooth optimization

AMS(MOS) subject classification. 15A99

1. Introduction.

In this paper we consider the problem of rotating m matrices towards a best least-squares fit. Let A_i, $i = 1, \cdots, m$, be a family of $n \times k$ matrices. Then it is necessary to find orthogonal $k \times k$ matrices Y_i, $i = 1, \cdots, m$, for which the function

$$f(Y_1, \cdots, Y_m) = \sum_{i<j} \operatorname{tr} (A_i Y_i - A_j Y_j)^T (A_i Y_i - A_j Y_j)$$

is minimized, or equivalently, for which the function

$$g(Y_1, \cdots, Y_m) = \sum_{i<j} \operatorname{tr} Y_i^T A_i^T A_j Y_j$$

is maximized. The problem has been discussed extensively in the psychometric literature and is known as the orthogonal Procrustes problem (see [2], [4], [8]–[10], and references therein). For $m = 2$ the solution is known and can be given in a closed form using the singular value decomposition of the matrix $A_1^T A_2$ (von Neumann [11]). That is, let $A_1^T A_2 = PDQ^T$, where P and Q are orthogonal matrices and D is a nonnegative definite diagonal matrix. Then $Y_1 = P$ and $Y_2 = Q$ solves the problem. It appears that the general case of $m > 2$ cannot be solved explicitly and an iterative procedure is required. A numerical algorithm employing singular value decompositions successively was proposed in Ten Berge [9]. It can be shown that this algorithm converges, but there is no guarantee that the calculated stationary point corresponds to the global optimum. Therefore various upper bounds for the maximum of the objective function g have been introduced [9].

We consider a dual approach to the Procrustes problem where the maximal value of the function g is approximated from above. The corresponding algorithm involves minimization of the sum of k largest eigenvalues of a symmetric matrix considered as a function of some elements of this matrix. The objective function is then convex, but is not necessarily differentiable. It will be shown that under conditions ensuring differentiability of the objective function at the minimum, this algorithm gives the global solution of the Procrustes problem.

* Received by the editors August 3, 1987; accepted for publication October 22, 1987.

† Department of Mathematics and Applied Mathematics, University of South Africa, P.O. Box 392, Pretoria 0001, South Africa.

2. Upper bounds. In this section we discuss some upper bounds for the maximal value of the function g. Consider the $k \times k$ matrices $S_{ij} = A_i^T A_j$ and let $S_{ij} = P_{ij} D_{ij} Q_{ij}^T$ be their singular value decompositions. Then

$$d^* = \sum_{i<j} \operatorname{tr} D_{ij} \tag{1}$$

gives an upper bound for the maximum of $g(Y_1, \cdots, Y_m)$ with respect to orthogonal matrices $Y_1, \cdots, Y_m$ [9, p. 273]. We consider another upper bound for the maximum of g. Let

$$S(X) = \begin{bmatrix} X_1 & S_{12} & \cdots & S_{1m} \\ S_{21} & X_2 & \cdots & S_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ S_{m1} & S_{m2} & \cdots & X_m \end{bmatrix}$$

be the $mk \times mk$ symmetric matrix considered as a function of the symmetric block diagonal matrix $X = \operatorname{diag}(X_1, \cdots, X_m)$. Denote by $\lambda_1(X) \geqq \cdots \geqq \lambda_{mk}(X)$ the eigenvalues of $S(X)$. Then the following result holds.

THEOREM 1. *For every X the number*

$$l(X) = \frac{1}{2}\left[m \sum_{i=1}^{k} \lambda_i(X) - \operatorname{tr} S(X)\right] \tag{2}$$

gives an upper bound for the maximum of g.

Proof. The sum of k largest eigenvalues of the symmetric matrix $S(X)$ can be represented in the form (Ky Fan [5])

$$\lambda_1(X) + \cdots + \lambda_k(X) = \max_Z \operatorname{tr} Z^T S(X) Z, \tag{3}$$

where the maximum in the right-hand side of (3) is taken over all $mk \times k$ matrices Z such that $Z^T Z = I_k$. (I_k denotes the $k \times k$ identity matrix.) Now let $Y_1, \cdots, Y_m$ be a set of orthogonal matrices and consider the corresponding $mk \times k$ matrix $Y = (Y_1^T, \cdots, Y_m^T)^T$. Then $Y^T Y = mI_k$, and hence it follows from (3) that

$$m \sum_{i=1}^{k} \lambda_i(X) \geqq \operatorname{tr} Y^T S(X) Y.$$

Moreover,

$$\operatorname{tr} Y^T S(X) Y = 2 \sum_{i<j} \operatorname{tr} Y_i^T S_{ij} Y_j + \sum_{i=1}^{m} \operatorname{tr} X_i,$$

and hence

$$g(Y_1, \cdots, Y_m) \leqq l(X).$$

Since the orthogonal matrices $Y_1, \cdots, Y_m$ are arbitrary, this completes the proof. □

It is natural now to minimize the function $l(X)$. We show in the next theorem that the obtained upper bound is always better than the upper bound given in (1).

THEOREM 2. *Let d^* be the upper bound given by the right-hand side of* (1). *Then*

$$d^* \geqq \inf l(X). \tag{4}$$

Proof. Let $P_{ij}D_{ij}Q_{ij}^T$ be singular value decompositions of the matrices S_{ij}. Consider the symmetric matrices

$$X_i^* = -\sum_{\substack{j=1 \\ j\neq i}}^{m} P_{ij}D_{ij}P_{ij}^T, \tag{5}$$

$i = 1, \cdots, m$. We show that

$$d^* \geqq l(X^*). \tag{6}$$

First we observe that

$$\operatorname{tr} S(X^*) = -\sum_{i\neq j} \operatorname{tr} D_{ij}.$$

Therefore in order to prove (6) it will be sufficient to show that the matrix $S(X^*)$ is nonpositive definite, i.e., $\lambda_1(X^*) \leqq 0$, and hence $\lambda_1(X^*) + \cdots + \lambda_k(X^*) \leqq 0$. Consider vectors $y_1, \cdots, y_m \in \mathbb{R}^k$. For any two vectors $a, b \in \mathbb{R}^k$ we have that

$$2a^Tb \leqq a^Ta + b^Tb,$$

and hence, by taking $a = D_{ij}^{1/2}P_{ij}^Ty_i$ and $b = D_{ij}^{1/2}Q_{ij}^Ty_j$, we obtain

$$y_i^TP_{ij}D_{ij}Q_{ij}^Ty_j \leqq \tfrac{1}{2}(y_i^TP_{ij}D_{ij}P_{ij}^Ty_i + y_j^TQ_{ij}D_{ij}Q_{ij}^Ty_j).$$

Now form the $mk \times 1$ vector $y = (y_1^T, \cdots, y_m^T)^T$. It then follows that

$$\begin{aligned} y^TS(X^*)y &= \sum_{i\neq j} y_i^TS_{ij}y_j + \sum_{i=1}^{m} y_i^TX_i^*y_i \\ &\leqq \frac{1}{2}\sum_{i\neq j} y_i^TP_{ij}D_{ij}P_{ij}^Ty_i + \frac{1}{2}\sum_{i\neq j} y_j^TQ_{ij}D_{ij}Q_{ij}^Ty_j + \sum_{i=1}^{m} y_i^TX_i^*y_i. \end{aligned}$$

Moreover, since $S_{ij}^T = S_{ji}$ we have that $Q_{ij}D_{ij}P_{ij}^T = P_{ji}D_{ji}Q_{ji}^T$, and hence we can choose $Q_{ij} = P_{ji}$. Therefore

$$\begin{aligned} y^TS(X^*)y &\leqq \sum_{i\neq j} y_i^TP_{ij}D_{ij}P_{ij}^Ty_i + \sum_i y_i^TX_i^*y_i \\ &= \sum_{i\neq j} y_i^TP_{ij}D_{ij}P_{ij}^Ty_i - \sum_i y_i^T\left(\sum_{j\neq i} P_{ij}D_{ij}P_{ij}^T\right)y_i = 0. \end{aligned}$$

Since vector y is arbitrary this proves that the matrix $S(X^*)$ is nonpositive definite, and hence the inequality (6) follows. Clearly (6) implies (4) and the proof is complete. □

Notice that the proof of Theorem 2 is constructive. Inequality (6) suggests X^* as a good starting point in minimization of the function $l(X)$. It was found in extensive numerical experimentations that this choice of the starting point is indeed a very good one.

Now we discuss some properties of the function $l(X)$. First we observe that it follows from the max-representation (3) that the sum $\sum_{i=1}^k \lambda_i(X)$, and hence the function $l(X)$, are convex. Using this max-representation it is also possible to calculate the subdifferential of $l(X)$ (cf. [1, Lemma 4.4]). For our purposes the following result will be particularly useful. Consider a block diagonal matrix X_0 and let $E = [e_1, \cdots, e_k]$ be an $mk \times k$ matrix whose columns $e_1, \cdots, e_k$ form a set of orthonormal eigenvectors of $S(X_0)$ corresponding to the eigenvalues $\lambda_1(X_0), \cdots, \lambda_k(X_0)$, respectively. We denote $\operatorname{diag}_X(EE^T)$

the block diagonal submatrix of EE^T corresponding to the block diagonal elements of X.

THEOREM 3. *The function $l(X)$ is the differentiable at X_0 if and only if $\lambda_k(X_0)$ is strictly greater than $\lambda_{k+1}(X_0)$. In the last case the gradient of l at X_0 is given by*

$$\nabla l(X_0) = \tfrac{1}{2}\{m \operatorname{diag}_X(EE^T) - I_{mk}\}. \tag{7}$$

Proof. Consider the max-representation (3). The function $\operatorname{tr} Z^T S(X)Z$ is linear in X and its gradient is given by $\operatorname{diag}_X(ZZ^T)$. Then it follows from a theorem of Danskin [3] that the subdifferential $\partial h(X)$ of the (convex) function $h(X) = \lambda_1(X) + \cdots + \lambda_k(X)$ is the convex hull of block diagonal matrices $\operatorname{diag}_X(ZZ^T)$ taken over all maximizers in the right-hand side of (3) (see Rockafellar [7, § 23] for the definition and basic properties of subdifferentials). Notice that a matrix Z is such a maximizer if and only if its columns form a set of orthonormal eigenvectors of $S(X)$ corresponding to k largest eigenvalues. It follows that the subdifferential $\partial h(X_0)$ is a singleton if and only if

$$\lambda_k(X_0) > \lambda_{k+1}(X_0). \tag{8}$$

Therefore the function $h(X)$ and then $l(X)$ are differentiable at X_0 if and only if (8) holds. In the last case (7) follows. □

Now let X_0 be a minimizer of $l(X)$ and consider the partition

$$E = [E_1^T, \cdots, E_m^T]^T,$$

E_i are $k \times k$, of the associated matrix E. Suppose that (8) holds. Then the gradient $\nabla l(X_0)$ is zero, and hence it follows from (7) that

$$m \operatorname{diag}_X(EE^T) = I_{mk}.$$

This implies that $mE_iE_i^T = I_k$, and hence

$$Y_i = m^{1/2}E_i, \qquad i = 1, \cdots, m \tag{9}$$

are orthogonal matrices. Moreover, the proof of Theorem 1 shows that in this case the orthogonal matrices Y_i given in (9) maximize the function g. We obtain that the dual problem of minimization of the function $l(X)$ not only provides an upper bound but actually *solves* the primary Procrustes problem if the corresponding minimizer X_0 satisfies (8).

3. Numerical experimentations. The main difficulty in numerical minimization of the function $l(X)$ is that it is not everywhere differentiable. Although considerable attention has been attracted to minimization of nondifferentiable convex functions, the developed algorithms are quite complicated and, what is more important, are slow to converge (cf. Cullum, Donath, and Wolfe [1]). In any case we are really interested in situations where the objective function $l(X)$ is differentiable at the minimum and consequently the dual problem provides a solution for the primary problem. Therefore some standard "differentiable" approaches have been applied. The point X^*, given in (5), proved to be a very good starting point. In fact, in many cases the gradient $\nabla l(X^*)$ was closed to zero so that the eigenvectors of $S(X^*)$, via formula (9), gave numerically acceptable solutions for the Procrustes problem.

A number of experiments were performed for various choices of m, n, and k. We briefly discuss two, namely $(m, n, k) = (8, 3, 3)$ and $(5, 4, 4)$. Denote by g_0 the maximal value of the function g as obtained by the Ten Berge algorithm proposed in [9] and by l_0 the minimal value of l as obtained by a slightly modified Newton method we will discuss later.

For each choice of (m, n, k) we generated ten sets, each containing m matrices $A_1, \cdots, A_m$. The entries of each A_i consisted of uniformly and independently distributed random numbers between -1 and 1. We first tried to minimize l by using the conjugate gradient algorithm. This turned out to be rather unsatisfactory since the convergence was slow, and in some cases it was not even achieved after 100 iterations.

Then the Newton method was tried. Notice that for computational purposes, the symmetry of X should be kept in mind. Therefore the function $l(X)$ was considered as a function of the $mk(k+1)/2 \times 1$ vector $x = (\text{vecs}^T X_1, \cdots, \text{vecs}^T X_m)^T$, where $\text{vecs} X_i$ stands for the $k(k+1)/2 \times 1$ symmetric mode vector representation of X_i. We denote by x_{ij} the entry of x corresponding to the (i, j) element of X. It follows that $\partial l/\partial x_{ii} = [\nabla l(X)]_{ii}$ and $\partial l/\partial x_{ij}$ is $2[\nabla l(X)]_{ij}$ when $i \neq j$. Here $\nabla l(X)$ is the gradient of l as it is given in (7). Now let $\{e_1, \cdots, e_{km}\}$ be a set of orthonormal eigenvectors of $S(X)$ corresponding to the eigenvalues $\lambda_1(X), \cdots, \lambda_{km}(X)$. Denote e_{ij} the ith entry of e_j. Then if we assume $\lambda_k(X) \neq \lambda_{k+1}(X)$ and let

$$a_{st,uv} = m \sum_{i=1}^{k} \sum_{j=k+1}^{km} (\lambda_i(X) - \lambda_j(X))^{-1} e_{si} e_{tj} e_{ui} e_{vj},$$

the elements of the $mk(k+1)/2 \times mk(k+1)/2$ Hessian matrix $H(X)$ of l at X can be shown to be

$$\frac{\partial^2 l}{\partial x_{st} \partial x_{uv}} = \begin{cases} a_{st,uv}, & s = t, \quad u = v, \\ a_{st,uv} + a_{ts,uv}, & s \neq t, \quad u = v, \\ a_{st,uv} + a_{st,vu}, & s = t, \quad u \neq v, \\ a_{st,uv} + a_{st,vu} + a_{ts,uv} + a_{ts,vu}, & s \neq t, \quad u \neq v \end{cases}$$

(cf. Lancaster [6]).

For each set of matrices $A_1, \cdots, A_m$ the value g_0 was calculated. Then the Newton algorithm for minimization of $l(X)$ was implemented with X^*, defined in Theorem 2, taken as the starting point. Notice that the Hessian matrix $H(X)$ is always *singular*. Therefore it was stabilized at each iteration by adding $\varepsilon = 0.1$ along its diagonal. After each iteration the new value of l was verified in order to see if it was less than the previous one. If not, a line search was performed along the direction $-H^{-1}(X)\nabla l(X)$. This became especially necessary after a number of iterations, when in some cases the kth and $(k+1)$th eigenvalues tend to converge to a common value. Convergence of the first algorithm was assumed when the difference between consecutive values of g was less than 0.0001. The difference value 0.01 was taken for the second. A point to which the Newton algorithm converged is denoted by X_0, i.e., $l_0 = l(X_0)$.

Tables 1 and 2 sum up the results obtained.

From these results the following observations are made:

(1) The point X^* seems to be a good starting point and $l(X^*)$ is a good approximation for l_0 and g_0.

(2) Whenever the difference $\lambda_k(X_0) - \lambda_{k+1}(X_0)$ is greater than about 0.1, the Newton algorithm converges in a few iterations and the corresponding value of $l_0 - g_0$ is very small. This suggests that the Ten Berge algorithm also converges in each such case to the global maximum of g.

(3) The kth and $(k+1)$th eigenvalues frequently converge to a common value, in which case the function l is not differentiable at the corresponding point X_0. Despite this the obtained upper bound l_0 is still a very good one.

(4) For all the data we generated the Ten Berge algorithm seemed to converge to the global maximum of g. This is rather surprising, since the corresponding problem is not a convex one and only certain stationarity of the calculated point is ensured by the

TABLE 1
Case $(m, n, k) = (8, 3, 3)$.

g_0	l_0	$l(X^*)$	$l_0 - g_0$	$l(X^*) - l_0$	$\lambda_k(X_0) - \lambda_{k+1}(X_0)$
83.074	83.072	83.095	−0.002	0.023	0.526136
71.997	71.996	72.126	−0.001	0.130	1.098260
63.835	63.864	64.079	0.029	0.215	0.000105
64.790	64.845	65.078	0.055	0.233	0.000706
64.102	64.102	64.837	0.000	0.735	0.000021
66.012	66.013	66.140	0.001	0.126	1.046920
65.310	65.317	65.670	0.007	0.353	0.000066
61.931	62.186	62.911	0.255	0.724	0.000061
73.106	73.346	73.829	0.240	0.483	0.000116
58.036	58.035	58.341	0.001	0.306	0.336850

TABLE 2
Case $(m, n, k) = (5, 4, 4)$.

g_0	l_0	$l(X^*)$	$l_0 - g_0$	$l(X^*) - l_0$	$\lambda_k(X_0) - \lambda_{k+1}(X_0)$
51.020	51.020	51.069	0.000	0.049	0.238905
46.016	46.078	46.199	0.061	0.121	0.000088
37.538	37.565	38.098	0.027	0.533	0.000109
44.478	44.524	44.708	0.047	0.183	0.000085
40.082	40.195	40.413	0.114	0.217	0.000115
36.503	36.546	36.880	0.044	0.333	0.000014
40.044	40.082	40.217	0.038	0.134	0.000964
45.232	45.428	45.942	0.196	0.514	0.000342
36.439	36.456	36.616	0.017	0.159	0.000986
45.284	45.436	46.623	0.152	1.187	0.002255

general theory. Whenever the difference $l_0 - g_0$ became reasonably significant, the difference $\lambda_k(X_0) - \lambda_{k+1}(X_0)$ was invariably very small and the Newton method did not converge.

REFERENCES

[1] J. CULLUM, W. E. DONATH, AND P. WOLFE, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Programming Stud., 3 (1975), pp. 35–55.

[2] N. CLIFF, *Orthogonal rotation to congruence*, Psychometrika, 31 (1966), pp. 33–42.

[3] J. M. DANSKIN, *The Theory of max-min and Its Applications to Weapons Allocation Problems*, Springer, New York, 1967.

[4] J. C. GOWER, *Generalized Procrustes analysis*, Psychometrika, 40 (1975), pp. 33–51.

[5] KY FAN, *On a theorem of Weyl concerning eigenvalues of linear transformations*, Proc. Nat. Acad. Sci. U.S.A., 35 (1949), pp. 652–655.

[6] P. LANCASTER, *On eigenvalues of matrices dependent on a parameter*, Numer. Math., 6 (1964), pp. 377–387.

[7] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

[8] P. H. SCHÖNEMANN, *A generalized solution of the orthogonal Procrustes problem*, Psychometrika, 31 (1966), pp. 1–10.

[9] J. M. F. TEN BERGE, *Orthogonal Procrustes rotation for two or more matrices*, Psychometrika, 42 (1977), pp. 267–276.

[10] J. M. F. TEN BERGE AND D. L. KNOL, *Orthogonal rotations to maximal agreement for two or more matrices of different column orders*, Psychometrika, 49 (1984), pp. 49–55.

[11] J. VON NEUMANN, *Some matrix inequalities and metrization of matrix space*, Tomsk Univ. Rev., 1 (1937), pp. 286–300.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

010

DETERMINANT INEQUALITIES VIA INFORMATION THEORY*

THOMAS M. COVER† AND JOY A. THOMAS‡

Abstract. Simple inequalities from information theory prove Hadamard's inequality and some of its generalizations. It is also proven that the determinant of a positive definite matrix is log-concave and that the ratio of the determinant of the matrix to the determinant of its principal minor $|K_n|/|K_{n-1}|$ is concave, establishing the concavity of minimum mean squared error in linear prediction. For Toeplitz matrices, the normalized determinant $|K_n|^{1/n}$ is shown to decrease with n.

Key words. inequalities, entropy, Hadamard, determinants

AMS(MOS) subject classification. 94A15

1. Introduction. The entropy inequalities of information theory have obvious intuitive meaning. For example, the entropy (or uncertainty) of a collection of random variables is less than or equal to the sum of their entropies. Letting the random variables be multivariate normal will yield Hadamard's inequality [1], [2]. We shall find many such determinant inequalities using this technique. We use throughout the fact that if

$$\phi_K(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|K|^{1/2}} e^{-(1/2)\mathbf{x}^t K^{-1}\mathbf{x}} \tag{1}$$

is the multivariate normal density with mean $\mathbf{0}$ and covariance matrix K, then the entropy $h(X_1, X_2, \cdots, X_n)$ is given by

$$h(X_1, X_2, \cdots, X_n) = -\int \phi_K \ln \phi_K = \frac{1}{2} \ln (2\pi e)^n |K|, \tag{2}$$

where $|K|$ denotes the determinant of K, and ln denotes the natural logarithm. This equality is verified by direct computation with the use of

$$\int \phi_K(\mathbf{x})\mathbf{x}^t K^{-1}\mathbf{x}\, d\mathbf{x} = \sum_i \sum_j K_{ij}(K^{-1})_{ij} = n = \ln e^n. \tag{3}$$

First we give some information theory preliminaries, then the determinant inequalities.

2. Information inequalities. In this section, we introduce some of the basic information theoretic quantities and prove a few simple inequalities using convexity. We assume throughout that the vector $(X_1, X_2, \cdots, X_n)$ has a probability density

$$f(x_1, x_2, \cdots, x_n).$$

We need the following definitions.

DEFINITION. The *entropy* $h(X_1, X_2, \cdots, X_n)$, sometimes written $h(f)$, is defined by

$$h(X_1, X_2, \cdots, X_n) = -\int f \ln f. \tag{4}$$

* Received by the editors October 28, 1987; accepted for publication November 2, 1987. This work was partially supported by Bell Communications Research and by National Science Foundation grant ECS85-20136. A preliminary version of this paper appears as Bellcore Technical Memo TM-ARH-010-203.

† Departments of Electrical Engineering and Statistics, Stanford University, Stanford, California 94305.

‡ Department of Electrical Engineering, Stanford University, Stanford, California 94305.

DEFINITION. The functional $D(f\|g) = \int f(\mathbf{x}) \ln (f(\mathbf{x})/g(\mathbf{x}))\, d\mathbf{x}$ is called the *relative entropy*, where f and g are probability densities.

The relative entropy $D(f\|g)$ is also known as the Kullback–Leibler information number, information for discrimination, and information distance. We also note that $D(f\|g)$ is the error exponent in the hypothesis test of f versus g.

DEFINITION. The *conditional entropy* $h(X\,|\,Y)$ of X, given Y, is defined by

$$h(X\,|\,Y) = -\int f(x,y) \ln f(x\,|\,y)\, dx\, dy. \tag{5}$$

We now observe certain natural properties of these information quantities.

LEMMA 1. $D(f\|g) \geqq 0$.

Proof. Let A be the support of f. Then by Jensen's inequality, $-D(f\|g) = \int_A f \ln (g/f) \leqq \ln \int_A f(g/f) = \ln \int_A g \leqq \ln 1 = 0$. □

LEMMA 2. *If (X, Y) have a joint density, then $h(X\,|\,Y) = h(X, Y) - h(Y)$.*

Proof. $h(X\,|\,Y) = -\int f(x, y) \ln f(x\,|\,y)\, dx\, dy = -\int f(x, y) \ln (f(x, y)/f(y))\, dx\, dy = -\int f(x, y) \ln f(x, y)\, dx\, dy + \int f(y) \ln f(y)\, dy = h(X, Y) - h(Y)$. □

LEMMA 3. *$h(X\,|\,Y) \leqq h(X)$, with equality if and only if X and Y are independent.*

Proof.

$$h(X) - h(X\,|\,Y) = \int f(x,y) \ln (f(x\,|\,y)/f(x)) = \int f(x,y) \ln (f(x,y)/f(x)f(y)) \geqq 0,$$

by $D(f(x, y)\|f(x)f(y)) \geqq 0$. Equality implies $f(x, y) = f(x)f(y)$ almost everywhere by strict concavity of the logarithm. □

LEMMA 4 (Chain Rule). *$h(X_1, X_2, \cdots, X_n) = \sum_{i=1}^n h(X_i|X_{i-1}, X_{i-2}, \cdots, X_1) \leqq \sum_{i=1}^n h(X_i)$ with equality if and only if $X_1, X_2, \cdots, X_n$ are independent.*

Proof. The equality is the chain rule for entropies, which we get by repeatedly applying Lemma 2. The inequality follows from Lemma 3, and we have equality if and only if $X_1, X_2, \cdots, X_n$ are independent. □

LEMMA 5. *If X and Y are independent, then $h(X + Y) \geqq h(X)$.*

Proof. $h(X + Y) \geqq h(X + Y\,|\,Y) = h(X\,|\,Y) = h(X)$. □

We will also need the entropy maximizing property of the multivariate normal.

LEMMA 6. *Let the random vector $\mathbf{X} \in \mathbf{R}^n$ have zero mean and covariance $K = E\mathbf{X}\mathbf{X}^t$, i.e., $K_{ij} = EX_iX_j$, $1 \leqq i, j \leqq n$. Then $h(\mathbf{X}) \leqq \frac{1}{2} \ln (2\pi e)^n |K|$, with equality if and only if $f(\mathbf{x}) = \phi_K(\mathbf{x})$.*

Proof. Let $g(\mathbf{x})$ be any density satisfying $\int g(\mathbf{x}) x_i x_j\, d\mathbf{x} = K_{ij}$, for all i, j. Then

$$\begin{aligned} 0 &\leqq D(g\|\phi_K) \\ &= \int g \ln (g/\phi_K) \\ &= -h(g) - \int g \ln \phi_K \qquad (6) \\ &= -h(g) - \int \phi_K \ln \phi_K \\ &= -h(g) + h(\phi_K), \end{aligned}$$

where the substitution $\int g \ln \phi_K = \int \phi_K \ln \phi_K$ follows from the fact that g and ϕ_K yield the same moments of the quadratic form $\ln \phi_K(\mathbf{x})$. □

Motivated by a desire to prove Szasz's generalization of Hadamard's inequality in the next section, we develop a new inequality on the entropy rates of random subsets of random variables. Let $(X_1, X_2, \cdots, X_n)$ have a density and for every $S \subseteq \{1, 2, \cdots, n\}$, denote by $X(S)$ the subset $\{X_i: i \in S\}$.

DEFINITION. Let

$$h_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:\,|S|=k} \frac{h(X(S))}{k}. \tag{7}$$

Here $h_k^{(n)}$ is the average entropy in bits per symbol of a randomly drawn k-element subset of $\{X_1, X_2, \cdots, X_n\}$. The following lemma states that the average entropy decreases monotonically in the size of the subset.

LEMMA 7.

$$h_1^{(n)} \geq h_2^{(n)} \geq \cdots \geq h_n^{(n)}. \tag{8}$$

Proof. We will first prove the last inequality, i.e., $h_n^{(n)} \leq h_{n-1}^{(n)}$. We write

$$h(X_1, X_2, \cdots, X_n) = h(X_1, X_2, \cdots, X_{n-1}) + h(X_n | X_1, X_2, \cdots, X_{n-1})$$

$$h(X_1, X_2, \cdots, X_n) = h(X_1, X_2, \cdots, X_{n-2}, X_n) + h(X_{n-1} | X_1, X_2, \cdots, X_{n-2}, X_n)$$

$$\leq h(X_1, X_2, \cdots, X_{n-2}, X_n) + h(X_{n-1} | X_1, X_2, \cdots, X_{n-2})$$

$$\vdots$$

$$h(X_1, X_2, \cdots, X_n) \leq h(X_2, X_3, \cdots, X_n) + h(X_1).$$

Adding these n inequalities and using the chain rule, we obtain

$$nh(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} h(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n) + h(X_1, X_2, \cdots, X_n) \tag{9}$$

or

$$\frac{1}{n} h(X_1, X_2, \cdots, X_n) \leq \frac{1}{n} \sum_{i=1}^{n} \frac{h(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)}{n-1}, \tag{10}$$

which is the desired result $h_n^{(n)} \leq h_{n-1}^{(n)}$.

We now prove that $h_k^{(n)} \leq h_{k-1}^{(n)}$ for all $k \leq n$, by first conditioning on a k-element subset, then taking a uniform choice over its $(k-1)$-element subsets. For each k-element subset, $h_k^{(k)} \leq h_{k-1}^{(k)}$, and hence the inequality remains true after taking the expectation over all k-element subsets chosen uniformly from the n elements. □

COROLLARY. *Let $r > 0$, and define*

$$g_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:\,|S|=k} e^{rh(X(S))/k}. \tag{11}$$

Then

$$g_1^{(n)} \geq g_2^{(n)} \geq \cdots \geq g_n^{(n)}. \tag{12}$$

Proof. Starting from (10) in the proof of Lemma 7, we multiply both sides by r, exponentiate, and then apply the arithmetic mean geometric mean inequality to obtain

(13)

$$\exp\left(\frac{1}{n} rh(X_1, X_2, \cdots, X_n)\right) \leqq \exp\left(\frac{1}{n}\sum_{i=1}^{n} \frac{rh(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)}{n-1}\right)$$

$$\leqq \frac{1}{n}\sum_{i=1}^{n} \exp\left(\frac{rh(X_1, X_2, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)}{n-1}\right)$$

for all $r \geqq 0$,

which is equivalent to $g_n^{(n)} \leqq g_{n-1}^{(n)}$. Now we use the same arguments as in Lemma 7, taking an average over all subsets to prove the result that for all $k \leqq n$, $g_k^{(n)} \leqq g_{k-1}^{(n)}$. □

Finally, we have the entropy power inequality, the only result we do not prove.

LEMMA 8. *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent random n-vectors with densities, then*

$$\exp\left(\frac{2}{n}h(\mathbf{X}+\mathbf{Y})\right) \geqq \exp\left(\frac{2}{n}h(\mathbf{X})\right) + \exp\left(\frac{2}{n}h(\mathbf{Y})\right). \tag{14}$$

Proof. See Shannon [3] for the statement and Stam [4] and Blachman [5] for the proof. Unlike the previous results, the proof is not elementary. □

3. Determinant inequalities. Throughout we will assume that K is a nonnegative definite symmetric $n \times n$ matrix. Let $|K|$ denote the determinant of K.

We first prove a result due to Ky Fan [6].

THEOREM 1. $\ln |K|$ *is concave.*

Proof. Let X_1 and X_2 be normally distributed n-vectors, $\mathbf{X}_i \sim \phi_{K_i}(\mathbf{x})$, $i = 1, 2$. Let the random variable θ have distribution $\Pr\{\theta = 1\} = \lambda$, $\Pr\{\theta = 2\} = 1 - \lambda$, $0 \leqq \lambda \leqq 1$. Let θ, $\mathbf{X}_1$, and $\mathbf{X}_2$ be independent and let $\mathbf{Z} = \mathbf{X}_\theta$. Then $\mathbf{Z}$ has covariance $K_Z = \lambda K_1 + (1 - \lambda)K_2$. However, $\mathbf{Z}$ will not be multivariate normal. By first using Lemma 6, followed by Lemma 3, we have

$$\begin{aligned} \tfrac{1}{2}\ln(2\pi e)^n|\lambda K_1 + (1-\lambda)K_2| &\geqq h(\mathbf{Z}) \geqq h(\mathbf{Z}|\theta) \\ &= \lambda\tfrac{1}{2}\ln(2\pi e)^n|K_1| + (1-\lambda)\tfrac{1}{2}\ln(2\pi e)^n|K_2|. \end{aligned} \tag{15}$$

Thus

$$|\lambda K_1 + (1-\lambda)K_2| \geqq |K_1|^{\lambda}|K_2|^{1-\lambda}, \tag{16}$$

as desired. □

The next theorem, used in [7], is too easy to require a new proof, but we provide it anyway.

THEOREM 2. $|K_1 + K_2| \geqq |K_1|$.

Proof. Let $\mathbf{X}$, $\mathbf{Y}$ be independent random vectors with $\mathbf{X} \sim \phi_{K_1}$ and $\mathbf{Y} \sim \phi_{K_2}$. Then $\mathbf{X} + \mathbf{Y} \sim \phi_{K_1+K_2}$ and hence $\frac{1}{2}\ln(2\pi e)^n|K_1 + K_2| = h(\mathbf{X} + \mathbf{Y}) \geqq h(\mathbf{X}) = \frac{1}{2}\ln(2\pi e)^n|K_1|$, by Lemma 5. □

We now give Hadamard's inequality using the proof in [2]. See also [1] for an alternative proof.

THEOREM 3 (Hadamard). $|K| \leqq \prod K_{ii}$, *with equality if and only if* $K_{ij} = 0$, $i \neq j$.

Proof. Let $\mathbf{X} \sim \phi_K$. Then

$$\frac{1}{2}\ln(2\pi e)^n|K| = h(X_1, X_2, \cdots, X_n) \leqq \sum h(X_i) = \sum_{i=1}^{n}\frac{1}{2}\ln 2\pi e|K_{ii}|, \tag{17}$$

with equality if and only if $X_1, X_2, \cdots, X_n$ are independent, i.e., $K_{ij} = 0$, $i \neq j$.

We now prove a generalization of Hadamard's inequality due to Szasz [9]. Let $K(i_1, i_2, \cdots, i_k)$ be the k-rowed principal submatrix of K formed by the rows and columns with indices $i_1, i_2, \cdots, i_k$.

THEOREM 4 (Szasz). *If K is a positive definite $n \times n$ matrix and P_k denotes the product of all the principal k-rowed minors of K, i.e.,*

$$P_k = \prod_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} |K(i_1, i_2, \cdots, i_k)|, \tag{18}$$

then

$$P_1 \geq P_2^{1/(n-1)} \geq P_3^{1/(n-2)} \geq \cdots \geq P_n. \tag{19}$$

Proof. Let $\mathbf{X} \sim \phi_K$. Then the theorem follows directly from Lemma 7, with the identification $h_k^{(n)} = (1/n) \ln P_k + \frac{1}{2} \ln 2\pi e$.

We can also prove a related theorem.

THEOREM 5. *Let K be a positive definite $n \times n$ matrix and let*

$$S_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} |K(i_1, i_2, \cdots, i_k)|^{1/k}. \tag{20}$$

Then

$$\frac{1}{n} \operatorname{tr}(K) = S_1^{(n)} \geq S_2^{(n)} \geq \cdots \geq S_n^{(n)} = |K|^{1/n}. \tag{21}$$

Proof. This follows directly from the corollary to Lemma 7, with the identification $g_k^{(n)} = (2\pi e) S_k^{(n)}$, and $r = 2$ in (11) and (12). □

We now prove a property of Toeplitz matrices, which are important as the covariance matrices of stationary random processes. A Toeplitz matrix K is characterized by the property that $K_{ij} = K_{rs}$ if $|i - j| = |r - s|$. Let K_k denote the principal minor $K(1, 2, \cdots, k)$. For such a matrix, the following property can be proved easily from the properties of the entropy function.

THEOREM 6. *If the positive definite $n \times n$ matrix K is Toeplitz, then*

$$|K_1| \geq |K_2|^{1/2} \geq \cdots \geq |K_{n-1}|^{1/(n-1)} \geq |K_n|^{1/n} \tag{22}$$

and $|K_k|/|K_{k-1}|$ is decreasing in k.

Proof. Let $(X_1, X_2, \cdots, X_n) \sim \phi_{K_n}$. Then the quantities $h(X_k | X_{k-1}, \cdots, X_1)$ are decreasing in k, since

$$\begin{aligned} h(X_k | X_{k-1}, \cdots, X_1) &= h(X_{k+1} | X_k, \cdots, X_2) \\ &\geq h(X_{k+1} | X_k, \cdots, X_2, X_1), \end{aligned} \tag{23}$$

where the equality follows from the Toeplitz assumption and the inequality from the fact that conditioning reduces entropy. Thus the running averages

$$\frac{1}{k} h(X_1, \cdots, X_k) = \frac{1}{k} \sum_{i=1}^{k} h(X_i | X_{i-1}, \cdots, X_1) \tag{24}$$

are decreasing in k. The theorem then follows from

$$h(X_1, X_2, \cdots, X_k) = \tfrac{1}{2} \ln (2\pi e)^k |K_k|. \qquad \square$$

Since $h(X_n|X_{n-1}, \cdots, X_1)$ is a decreasing sequence, it has a limit. Hence by the Cesáro Mean Theorem,

$$\lim_{n\to\infty} \frac{h(X_1, X_2, \cdots, X_n)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} h(X_k|X_{k-1}, \cdots, X_1) = \lim_{n\to\infty} h(X_n|X_{n-1}, \cdots, X_1). \tag{25}$$

Translating this to determinants, we obtain the following result:

$$\lim_{n\to\infty} |K_n|^{1/n} = \lim_{n\to\infty} \frac{|K_n|}{|K_{n-1}|}, \tag{26}$$

which is one of the simple limit theorems for determinants that can be proved using information theory.

In problems connected with maximum entropy spectrum estimation, we would like to maximize the value of the determinant of a Toeplitz matrix, subject to constraints on the values in a band around the main diagonal. Choi and Cover [10] use information theoretic arguments to show that the matrix maximizing the determinant under these constraints is the Yule–Walker extension of the values along the band.

The proof of the next inequality (Oppenheim [11], Marshall and Olkin [12, p. 475]) follows immediately from the entropy power inequality, but because of the complexity of the proof of the entropy power inequality, is not offered as a simpler proof.

THEOREM 7 (Minkowski inequality [13]).

$$|K_1 + K_2|^{1/n} \geqq |K_1|^{1/n} + |K_2|^{1/n}. \tag{27}$$

Proof. Let $\mathbf{X}_1$, $\mathbf{X}_2$ be independent with $\mathbf{X}_i \sim \phi_{K_i}$. Noting that $\mathbf{X}_1 + \mathbf{X}_2 \sim \phi_{K_1+K_2}$ and using the entropy power inequality (Lemma 8) yields

$$\begin{aligned}(2\pi e)|K_1 + K_2|^{1/n} &= e^{(2/n)h(\mathbf{X}_1+\mathbf{X}_2)} \\ &\geqq e^{(2/n)h(\mathbf{X}_1)} + e^{(2/n)h(\mathbf{X}_2)} \\ &= (2\pi e)|K_1|^{1/n} + (2\pi e)|K_2|^{1/n}.\end{aligned} \tag{28}$$

□

4. Inequalities for ratios of determinants. We first prove a stronger version of Hadamard's theorem due to Ky Fan [8].

THEOREM 8. *For all* $1 \leqq p \leqq n$,

$$\frac{|K|}{|K(p+1, p+2, \cdots, n)|} \leqq \prod_{i=1}^{p} \frac{|K(i, p+1, p+2, \cdots, n)|}{|K(p+1, p+2, \cdots, n)|}. \tag{29}$$

Proof. We use the same idea as in Theorem 3, except that we use the conditional form of Lemma 3:

$$\begin{aligned}\frac{1}{2} \ln (2\pi e)^p \frac{|K|}{|K(p+1, p+2, \cdots, n)|} &= h(X_1, X_2, \cdots, X_p|X_{p+1}, X_{p+2}, \cdots, X_n) \\ &\leqq \sum h(X_i|X_{p+1}, X_{p+2}, \cdots, X_n) \\ &= \sum_{i=1}^{p} \frac{1}{2} \ln 2\pi e \frac{|K(i, p+1, p+2, \cdots, n)|}{|K(p+1, p+2, \cdots, n)|}.\end{aligned} \tag{30}$$

□

Before developing Theorem 9, we make an observation about minimum mean squared error linear prediction. If $(X_1, X_2, \cdots, X_n) \sim \phi_{K_n}$, we know that the conditional density of X_n given $(X_1, X_2, \cdots, X_{n-1})$ is univariate normal with mean linear in $X_1, X_2, \cdots, X_{n-1}$ and conditional variance σ_n^2. Here σ_n^2 is the minimum mean squared error $E(X_n - \hat{X}_n)^2$ over all linear estimators $\hat{X}_n$ based on $X_1, X_2, \cdots, X_{n-1}$.

LEMMA 9. $\sigma_n^2 = |K_n|/|K_{n-1}|$.

Proof. Using the conditional normality of X_n, Lemma 2 results in

$$\begin{aligned} \tfrac{1}{2} \ln 2\pi e\sigma_n^2 &= h(X_n|X_1, X_2, \cdots, X_{n-1}) \\ &= h(X_1, X_2, \cdots, X_n) - h(X_1, X_2, \cdots, X_{n-1}) \\ &= \tfrac{1}{2} \ln (2\pi e)^n |K_n| - \tfrac{1}{2} \ln (2\pi e)^{n-1} |K_{n-1}| \\ &= \tfrac{1}{2} \ln 2\pi e |K_n|/|K_{n-1}|. \end{aligned} \tag{31}$$

□

Minimization of σ_n^2 over a set of allowed covariance matrices $\{K_n\}$ is aided by the following theorem.

THEOREM 9. $\ln (|K_n|/|K_{n-p}|)$ *is concave in* K_n.

Proof. We remark that Theorem 1 cannot be used because $\ln (|K_n|/|K_{n-p}|)$ is the difference of two concave functions. Let $\mathbf{Z} = \mathbf{X}_\theta$, where $\mathbf{X}_1 \sim \phi_{S_n}(\mathbf{x})$, $\mathbf{X}_2 \sim \phi_{T_n}(\mathbf{x})$, $\Pr\{\theta = 1\} = \lambda = 1 - \Pr\{\theta = 2\}$, and $\mathbf{X}_1, \mathbf{X}_2, \theta$ are independent. The covariance matrix K_n of $\mathbf{Z}$ is given by

$$K_n = \lambda S_n + (1 - \lambda) T_n. \tag{32}$$

The following chain of inequalities proves the theorem:

$$\begin{aligned} &\lambda \frac{1}{2} \ln (2\pi e)^p |S_n|/|S_{n-p}| + (1-\lambda) \frac{1}{2} \ln (2\pi e)^p |T_n|/|T_{n-p}| \\ &\quad \overset{(a)}{=} \lambda h(X_{1n}, X_{1,n-1}, \cdots, X_{1,n-p+1}|X_{11}, \cdots, X_{1,n-p}) \\ &\qquad + (1-\lambda) h(X_{2n}, X_{2,n-1}, \cdots, \\ &\qquad\qquad X_{2,n-p+1}|X_{21}, \cdots, X_{2,n-p}) \\ &\quad = h(Z_n, Z_{n-1}, \cdots, Z_{n-p+1}|Z_1, \cdots, Z_{n-p}, \theta) \\ &\quad \overset{(b)}{\leqq} h(Z_n, Z_{n-1}, \cdots, Z_{n-p+1}|Z_1, \cdots, Z_{n-p}) \\ &\quad \overset{(c)}{\leqq} \frac{1}{2} \ln (2\pi e)^p \frac{|K_n|}{|K_{n-p}|}, \end{aligned} \tag{33}$$

where (a) follows from

$$h(X_n, X_{n-1}, \cdots, X_{n-p+1}|X_1, \cdots, X_{n-p}) = h(X_1, \cdots, X_n) - h(X_1, \cdots, X_{n-p}),$$

(b) follows from the conditioning lemma, and (c) follows from a conditional version of Lemma 6. □

The above theorem for the case $p = 1$ is due to Bergstrøm [14]. However, for $p = 1$, we can prove an even stronger theorem, also due to Bergstrøm [14].

THEOREM 10. $|K_n|/|K_{n-1}|$ *is concave in* K_n.

Proof. Again we use the properties of Gaussian random variables. Let us assume that we have two independent Gaussian random vectors, $\mathbf{X} \sim \phi_{A_n}$ and $\mathbf{Y} \sim \phi_{B_n}$. Let

$\mathbf{Z} = \mathbf{X} + \mathbf{Y}$. Then

$$\begin{aligned}
\frac{1}{2}\ln 2\pi e\frac{|A_n+B_n|}{|A_{n-1}+B_{n-1}|} &\overset{(a)}{=} h(Z_n|Z_{n-1},Z_{n-2},\cdots,Z_1)\\
&\overset{(b)}{\geqq} h(Z_n|Z_{n-1},Z_{n-2},\cdots,Z_1,X_{n-1},X_{n-2},\cdots,X_1,\\
&\qquad\qquad Y_{n-1},Y_{n-2},\cdots,Y_1)\\
&\overset{(c)}{=} h(X_n+Y_n|X_{n-1},X_{n-2},\cdots,X_1,Y_{n-1},Y_{n-2},\cdots,Y_1)\\
&\overset{(d)}{=} E\frac{1}{2}\ln\,[2\pi e\,\mathrm{Var}\,(X_n+Y_n|X_{n-1},X_{n-2},\cdots,X_1,\\
&\qquad\qquad Y_{n-1},Y_{n-2},\cdots,Y_1)] \qquad (34)\\
&\overset{(e)}{=} E\frac{1}{2}\ln\,[2\pi e(\mathrm{Var}\,(X_n|X_{n-1},X_{n-2},\cdots,X_1)\\
&\qquad\qquad +\mathrm{Var}\,(Y_n|Y_{n-1},Y_{n-2},\cdots,Y_1))]\\
&\overset{(f)}{=} E\frac{1}{2}\ln\left(2\pi e\left(\frac{|A_n|}{|A_{n-1}|}+\frac{|B_n|}{|B_{n-1}|}\right)\right)\\
&= \frac{1}{2}\ln\left(2\pi e\left(\frac{|A_n|}{|A_{n-1}|}+\frac{|B_n|}{|B_{n-1}|}\right)\right).
\end{aligned}$$

In the above derivation, (a) follows from Lemma 9, (b) from the fact the conditioning decreases entropy, and (c) from the fact that Z is a function of X and Y. $X_n + Y_n$ is Gaussian conditioned on $X_1, X_2, \cdots, X_{n-1}, Y_1, Y_2, \cdots, Y_{n-1}$, and hence we can express its entropy in terms of its variance, obtaining (d). Then (e) follows from the independence of X_n and Y_n conditioned on the past $X_1, X_2, \cdots, X_{n-1}, Y_1, Y_2, \cdots, Y_{n-1}$, and (f) follows from the fact that for a set of jointly Gaussian random variables, the conditional variance is constant, independent of the conditioning variables (Lemma 9). In general, by setting $A = \lambda S$ and $B = \bar{\lambda} T$, we obtain

$$\frac{|\lambda S_n+\bar{\lambda}T_n|}{|\lambda S_{n-1}+\bar{\lambda}T_{n-1}|} \geqq \lambda\frac{|S_n|}{|S_{n-1}|}+\bar{\lambda}\frac{|T_n|}{|T_{n-1}|}, \qquad (35)$$

i.e., $|K_n|/|K_{n-1}|$ is concave. Simple examples show that $|K_n|/|K_{n-p}|$ is not necessarily concave for $p \geqq 2$. □

5. Remarks. Concavity and Jensen's inequality play a role in all the proofs. The inequality $D(f\|g) = \int f \ln (f/g) \geqq 0$ is at the root of most of them.

Acknowledgment. We thank A. Bernardi and S. Pombra for their contributions.

REFERENCES

[1] A. MARSHALL AND I. OLKIN, *A convexity proof of Hadamard's inequality*, Amer. Math. Monthly, 89 (1982), pp. 687–688.

[2] T. COVER AND A. EL GAMAL, *An information theoretic proof of Hadamard's inequality*, IEEE Trans. Inform. Theory, IT-29 (1983), pp. 930–931.

[3] C. E. SHANNON, *A mathematical theory of communication*, Bell System Tech. J., 27 (1948), pp. 623–656.
[4] A. STAM, *Some inequalities satisfied by the quantities of information of Fisher and Shannon*, Inform. and Control, 2 (1959), pp. 101–112.
[5] N. BLACHMAN, *The convolution inequality for entropy powers*, IEEE Trans. Inform. Theory, IT-11 (1965), pp. 267–271.
[6] KY FAN, *On a theorem of Weyl concerning the eigenvalues of linear transformations*, II, Proc. Nat. Acad. Sci. U.S.A., 36 (1950), pp. 31–35.
[7] S. POMBRA AND T. COVER, *Gaussian Feedback Capacity*, Technical Report 63, Dept. of Statistics, Stanford Univ., Stanford, CA, September 1987; IEEE Trans. Inform. Theory, to appear.
[8] KY FAN, *Some inequalities concerning positive-definite matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 414–421.
[9] L. MIRSKY, *On a generalization of Hadamard's determinantal inequality due to Szasz*, Arch. Math., VIII (1957), pp. 274–275.
[10] B. S. CHOI AND T. M. COVER, *An information-theoretic proof of Burg's maximum entropy spectrum*, Proc. IEEE, 72 (1984), pp. 1094–1095.
[11] A. OPPENHEIM, *Inequalities connected with definite Hermitian forms*, J. London Math. Soc., 5 (1930), pp. 114–119.
[12] A. MARSHALL AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.
[13] H. MINKOWSKI, *Diskontinuitätsbereich für arithmetische Äquivalenz*, Journal für Math., 129 (1950), pp. 220–274.
[14] R. BELLMAN, *Notes on matrix theory—IV: an inequality due to Bergstrøm*, Amer. Math. Monthly, 62 (1955), pp. 172–173.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

011

TOMOGRAPHY IN PROJECTIVE SPACES: A HEURISTIC FOR LIMITED ANGLE RECONSTRUCTIVE MODELS*

PABLO M. SALZBERG†

Abstract. Given a projective geometry, **PG**(s, p), over a finite field of characteristic p, the tomographic problem can be stated in the following terms: unknown densities $w_1, w_2, \cdots$, are attached to each point of the space; how can these values be obtained if we are allowed to "irradiate" along the lines of our geometry and measure the total attenuation of energy of each ray (line)?

A complete answer to this problem is given. The solution is close to the (Filtered) Back Projection Technique, one of the earliest techniques in computerized tomography, and can be used as a spatial limited angle tomographic model.

Key words. tomography, mathematical modelling, image processing, matrix equations, latin squares, finite geometry

AMS(MOS) subject classifications. 68U10, 15A24, 05B15, 51E20

1. Preliminaries. The tomographic problem ([2], [6], [8]–[10]) can be succinctly described in the following terms: let **E** be a finite set endowed with a set **R** of "rays" or "directions" ($\mathbf{R} \subset 2^{\mathbf{E}}$), and let us assume that unknown densities $w_1, w_2, \cdots$, are assigned to the points of **E**. We would like to know the value of these densities, for which we are allowed to "irradiate" along the rays on **E** and measure the total attenuation of energy in each direction. Actually, we shall deal with the following equivalent problem: unknown weights $w_1, w_2, \cdots$, are attached to each point of **E**; how can these values be obtained if we know the weight of each ray in **R**?

Some discrete models were extensively studied in the earlier days of X-ray tomography (cf. [1] and references therein). In general, the tomographic problem involved in these models can be described in terms of a pair (**E**, **R**); **E** being a square grid (or matrix) of points in the Euclidean plane, whereas each ray in **R** consists of the subset of points belonging to a straight strip in some direction. Thus, the solution of the tomographic problem leads to the inversion of a large matrix which, besides the amount of computation required, is an ill-conditioned problem. One of the simplest techniques used in computerized tomography that avoids inverting large matrices is the Back Projection Technique (BPT) introduced by Kuhl and Edwards [5]. Unfortunately, as we shall see later, BPT combines the information obtained from the scanning process in an inadequate way, reconstructing the "images" (densities) without the resolving power of other methods. On the other hand, BPT is less sensitive to noise in data, which is one of the major problems when dealing with image reconstruction.

In what follows we shall exhibit a technique that is close to BPT arising from properties of rays in projective spaces. From this framework, it becomes clear why BPT fails to give better image resolution.

2. The PSCT model. As we mentioned above, a fairly simple solution to the tomographic problem can be found if $\mathbf{E} = \mathbf{PG}(s, p)$, the projective space of dimension s on $\mathbf{K}_p$ [3], [4], and **R** is the set of lines on **E**. In general, the possibility of finding a

* Received by the editors October 10, 1986; accepted for publication (in revised form) November 2, 1987. This work was supported by the National Institutes of Health (Minority Biomedical Research Support Program) under grant 1 S14 RRO3232-01.

† Department of Natural Sciences, University of the Sacred Heart, Santurce, Puerto Rico 00914.

solution depends on the properties of **R**. We shall consider the following four properties of projective lines:

(i) Given two different points P, Q there is one and only one line containing (incident to) P and Q.

(ii) Given two different points P, Q there is at least one line passing through P and not containing Q.

(iii) Two different lines are disjoint (parallel) or intersect at one point.

(iv) Every line has the same number of points.

Properties (i)–(iv) are not exclusive of projective spaces. Indeed, given any finite set **E**, let $\mathbf{R} = \{(P, Q)\colon P, Q \in \mathbf{E}\}$. Then it is clear that these four properties are satisfied.

Let us assume, momentarily, that **E** is any finite set of elements endowed with a set $\mathbf{R} = \{r_i\}_{i \in I}$, satisfying properties (i)–(iv). Given any point $P \in \mathbf{E}$, let $r_0^P, r_1^P, \cdots, r_t^P$ be the pencil of lines passing through P. Then as straightfoward consequences of these four properties we have the following equalities:

$$\text{(I)} \qquad \bigcap_{i=0}^{t} r_i^P = \{P\},$$

$$\text{(II)} \qquad \bigcup_{i=0}^{t} r_i^P = \mathbf{E}.$$

Furthermore, if we assume that each line has, say, $n + 1$ points (cf. (iv)), then it can be easily seen from (I) and (II) that the cardinality of **E** is $n(t + 1) + 1$; hence, *the number* $(t + 1)$ *of lines incident to any point remains constant.*

Now, to find the density w_P associated with any point $P \in \mathbf{E}$, let us denote by S_r the total weight of line r, i.e., $S_r = \sum_{Q \in r} w_Q$, and let S_i, for $i = 0, \cdots, t$, be the weights of the lines passing through P.

From (I) and (II) it is clear that the following equality holds:

$$\sum_{i=0}^{t} S_i = t w_P + T \tag{1}$$

where $T = \sum_{P \in \mathbf{E}} w_P$ is the total weight of the space **E**, a parameter which can be easily obtained from the weights of the rays. (Indeed, by adding both terms of (1) over all **E**, we obtain $T = (1/(t + 1)) \sum_{r \in \mathbf{R}} S_r$.)

Hence, solving (1) for w_P yields

$$w_P = (1/t)\left(\sum_{i=0}^{t} S_i\right) - (1/t)T. \tag{2}$$

Therefore, (2) furnishes a *fast inversion formula* for evaluating the unknown weights assigned to each point of **E**. It is also very close to that used in BPT [5]. The main difference resides precisely in the notion of ray itself. Indeed, lines in our projective geometry are difficult objects from those of the s-dimensional Euclidean space (see Fig. 2 in § 3).

3. An application: limited angle planar computerized tomography. In this section we develop, from the above theoretical framework of tomography in a projective space (PSCT), the heuristic for an applied technique.

Given a projective space $\mathbf{PG}(2, p)$, we can construct a complete set $\mathbf{M}^1, \cdots, \mathbf{M}^{p-1}$ of orthogonal latin squares [3], [7]. This complete set of such $p \times p$ matrices $\mathbf{M}^k = [m_{ij}^k]$ can also be obtained by means of the expressions $m_{ij}^k = i + k*j$, where

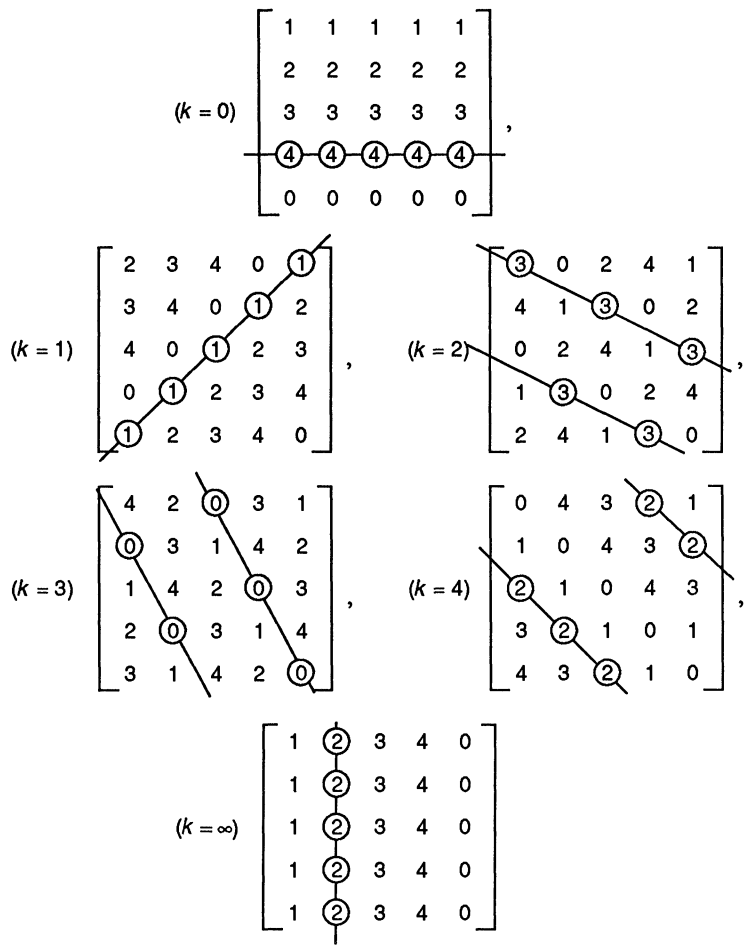


FIG. 1

$1 \leqq k \leqq p - 1$, $1 \leqq i, j \leqq p$, and "+" and "$*$" denote the operations on the field **K**. We shall add to the set the "canonical squares" $\mathbf{M}^0 = [m_{ij}^0]$ and $M^p = [m_{ij}^p]$, where $m_{ij}^0 = i$ for all j, and $m_{ij}^p = j$ for all i, which will be useful for introducing lines of slope 0 and ∞, respectively. Figure 1 shows a complete set of orthogonal squares for $p = 5$.

In this context, given a set **E** consisting of p^2 points arranged as a $p \times p$ square, each matrix $\mathbf{M}^k$ will define on **E** the set of parallel lines having slope k. More specifically, if we overlap the square $\mathbf{M}^k$ on **E**, a ray is defined as the set of those points of **E** receiving the same value. The set of points encircled in each square of Fig. 1 determines a line with the given slope k, once this matrix is overlapped on **E**. The whole set of encircled lines constitutes the pencil through allocation (4, 2).

Now, Fig. 2 illustrates the difference between lines in BPT and PSCT.

We now show how to apply the preceding theory in "reconstructing" a square.

Example. Let **M** be the matrix shown in Fig. 3, and assume it is a "black box" of which we can only know the sum of the assigned values along some directions.

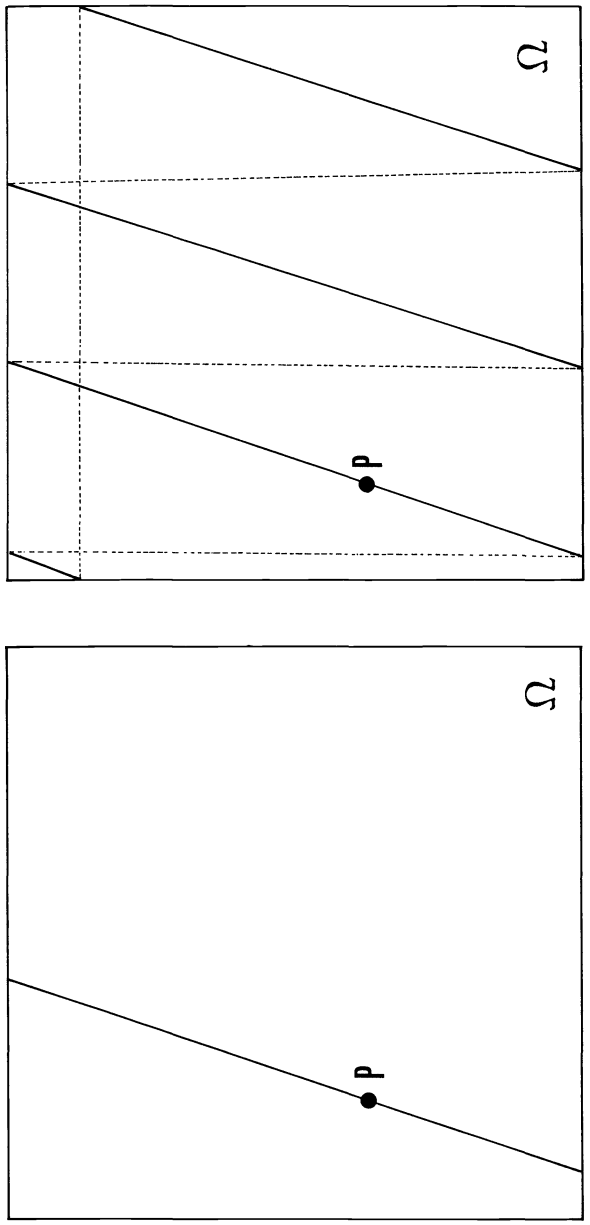


FIG. 2

$$\begin{bmatrix} 0 & -3 & 2 & -1 & 4 \\ 1 & 5 & 0 & -2 & 4 \\ 3 & 1 & -3 & 4 & 0 \\ 0 & 4 & -1 & 3 & -2 \\ 5 & 5 & 5 & 5 & 5 \end{bmatrix}$$

FIG. 3

TABLE 1

Slope	0	1	2	3	4	∞
Accumulated values ("attenuation")	4	8	9	16	15	12

In order to find out, for instance, the value assigned to location (4, 2) (which in this case is 4) we consider the pencil passing through (4, 2). These lines are precisely those encircled in Fig. 1. Thus, when scanning along each of these lines we find the values exhibited in Table 1.

Hence, according to formula (2),

$$w_{(4,2)} = (\tfrac{1}{5})(4+8+9+16+15+12) - (\tfrac{1}{5})T.$$

By scanning along a complete set of parallel lines (and adding the values) we obtain $T = 44$. Thus, $w_{(4,2)} = (\frac{1}{5})64 - (\frac{1}{5})44 = 4$.

Finally, in testing this planar CT model, the scanning field can be considered an $n \times n$ matrix whose entries are the unknown densities. For all practical purposes this matrix will represent a grid overlapped over the real object, whose norm is small enough so that we can assume its elements are of uniform density. In a computer screen, each element of this grid will consist of a few pixels.

Once we assign densities to the elements of the grid, i.e., once we have a mathematical phantom, we proceed to reconstruct the image by scanning with beams of p parallel bands ($p \leqq n$, p the power of a prime) irradiated in each of the $p + 1$ directions. To reconstruct the image, the preceding theory yields a matrix consisting of $p \times p$ "points," each of uniform density.

When we deal with this reconstruction procedure, a very interesting problem arises, namely, *to determine the pattern of crosses of Euclidean lines when overlapped to projective lines.* This is an open problem.

REFERENCES

[1] R. GORDON, *A tutorial on* ART (*Algebraic Reconstruction Techniques*), IEEE Trans. Nuclear Science, 21 (1974), pp. 78–93.

[2] J. F. GREENLEAF, *Computerized tomography with ultrasound*, Proc. IEEE, 71 (1983), pp. 330–337.

[3] M. HALL, JR., *Combinatorial Theory*, Blaisdell, Waltham, MA, 1967.

[4] J. W. P. HIRSCHFELD, *Projective Geometries Over Finite Fields*, Clarendon Press, Oxford, 1979.

[5] D. E. KUHL AND R. Q. EDWARDS, *Reorganizing data from transverse section scans of the brain using digital processing*, Radiology, 91 (1968), pp. 975–983.
[6] F. NATTERER, *The Mathematics of Computerized Tomography*, John Wiley, New York, 1985.
[7] D. RAGHAVARAO, *Constructions and Combinatorial Problems in Design of Experiments*, John Wiley, New York, 1971.
[8] L. A. SHEPP AND J. B. KRUSKAL, *Computerized tomography: the new medical X-ray technology*, Amer. Math. Monthly, 85 (1978), pp. 420–439.
[9] L. A. SHEPP, *Computerized tomography and nuclear magnetic resonance*, J. Computer Assisted Tomography, 4 (1980), pp. 94–107.
[10] M. M. TER-POGOSSIAN, M. E. RAICHLE, AND B. E. SOBEL, *Positron emission tomography*, Scientific American, 243 (1980), pp. 170–181.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

012

EIGENVECTORS OF DISTANCE-REGULAR GRAPHS*

DAVID L. POWERS†

Abstract. The objective of this work is to find properties of a distance-regular graph G that are expressed in the eigenvectors of its adjacency matrix. The approach is to consider the rows of a matrix of orthogonal eigencolumns as (coordinates of) points in Euclidean space, each one corresponding to a vertex of G. For the second eigenvalue, the symmetry group of the points is isomorphic to the automorphism group of G. Adjacency of vertices is related to linear dependence, linear independence, and proximity of points. Relative position of points is studied by way of the polytope that is their convex hull. Several families of examples are included.

Key words. eigenvector, distance-regular graph

AMS(MOS) subject classifications. 05C50, 15A18

1. Introduction. In this paper, we find properties of a distance-regular graph that are reflected in properties of the eigenvectors of the adjacency matrix. This class is chosen because of convenient algebraic properties. In a recent paper [9], C. Godsil uses similar ideas to bound the diameter of a distance-regular graph in terms of eigenvalue multiplicity. To some extent, the techniques used here were inspired by those of Terwilliger [14]. In this Introduction, we present definitions and a key lemma. Throughout, the graph G is assumed to have vertex set $V = \{1, 2, \cdots, n\}$.

DEFINITION. A partition of the vertex set of a graph G into $V_1, V_2, \cdots, V_t$ is a *coloration* if, for each i and j, each vertex h in V_i is adjacent to the same number, b_{ij}, of vertices in V_j. The square matrix $B = [b_{ij}]$ is called the *coloration matrix* of the partition.

A matrix-theoretic definition is this. Let X be the incidence matrix of the partition; that is, X has a one in the i,j-position if vertex i is in V_j or a zero otherwise. Then the partition is a coloration if and only if $AX = XB$ for some B, where A is the adjacency matrix of the graph. If the condition is fulfilled, then B is in fact the coloration matrix. Note that every row of X is a row of the identity, and no column is empty; thus X has independent columns. A general reference for colorations is [5, Chap. 4].

DEFINITION. A graph G is *distance-regular* if, for each vertex i of G, the *distance partition starting at i*,

$$V_0 = \{i\}, \qquad V_k = \{j : \text{dist}\,(i,j) = k\}$$

is a coloration, and the coloration matrix B is independent of i.

The coloration matrix B of a distance partition in a distance-regular graph has $d+1$ rows and columns, where d is the diameter of G. The triangle inequality in the graph guarantees that B is tridiagonal, and the implied connectedness makes the entries next to the diagonal nonzero. Furthermore, a distance-regular graph must be regular, of valence ρ, say, and the sum of the entries in each row of B is ρ. We follow the notation of [2] for the elements of B (although the matrix called B there is the transpose of this):

$$B = \begin{bmatrix} 0 & b_0 & 0 & 0 & \cdots & & & & & \\ c_1 & a_1 & b_1 & 0 & 0 & \cdots & & & & \\ 0 & c_2 & a_2 & b_2 & 0 & 0 & \cdots & & & \\ & & \cdots & & & & & & & \\ 0 & 0 & \cdots & & 0 & 0 & \cdots & 0 & c_d & a_d \end{bmatrix}.$$

* Received by the editors July 1, 1987; accepted for publication (in revised form) December 4, 1987. This work was supported by the Office of Naval Research under grant N00014-85-K-04097.

† Department of Mathematics and Computer Science, Clarkson University, Potsdam, New York 13676.

A familiar example of a distance-regular graph is the skeleton of a cube shown in Fig. 1. The coloration matrix of the distance partition is

$$B = \begin{bmatrix} 0 & 3 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 3 & 0 \end{bmatrix}.$$

If G is any graph on n vertices, the *distance matrices* $A_0, A_1, \cdots, A_d$, where d is the diameter of G, are defined by the requirement that A_k have a one in the i, j-position if dist $(i, j) = k$ and a zero otherwise. Obviously, $A_0 = I$ and $A_1 = A$, the adjacency matrix. The following theorem about distance-regular graphs is well known (see [2, p. 140]).

THEOREM A. *If G is a distance-regular graph, then the distance matrices $A_0, A_1, \cdots, A_d$ form a basis for the algebra of polynomials in the adjacency matrix $A = A_1$.*

For any symmetric $n \times n$ matrix A, if α is an eigenvalue of multiplicity m, there is an $n \times m$ matrix Z satisfying $AZ = \alpha Z$, $Z^T Z = I_m$. We call such a matrix, composed of orthonormal eigencolumns associated with α, a *complete eigenmatrix.* Then the projector associated with α is $L = ZZ^T$, which satisfies (see, e.g., [10, p. 196]) $AL = \alpha L$, $L^2 = L$, and rank $(L) = m$. If A is the adjacency matrix of a distance-regular graph, Theorem A guarantees that there are coefficients $y_0, y_1, \cdots, y_d$, such that

$$L = y_0 A_0 + y_1 A_1 + \cdots + y_d A_d, \tag{1}$$

because L is a polynomial in A. As the following lemma shows, the coefficients have a further significance in this case. This lemma is implicit in [2, Thm. 21.4, p. 143].

LEMMA 1. *Let A be the adjacency matrix of a distance-regular graph G, let α be an eigenvalue with multiplicity m, and let L be the associated projector. Then α is an eigenvalue of the coloration matrix B, and the column matrix $y = [y_0, y_1, \cdots, y_d]^T$, containing the coefficients from* (1), *satisfies $By = \alpha y$ and $y_0 = m/n$.*

Proof. Let X be the incidence matrix of the distance partition starting at vertex i, and consider column i of L (e_i is column i of an identity matrix whose dimension is dictated by context):

$$\begin{aligned} Le_i &= y_0 e_i + y_1 A e_i + \cdots + y_d A_d e_i \\ &= y_0 X e_1 + y_1 X e_2 + \cdots + y_d X e_{d+1} \\ &= Xy. \end{aligned}$$

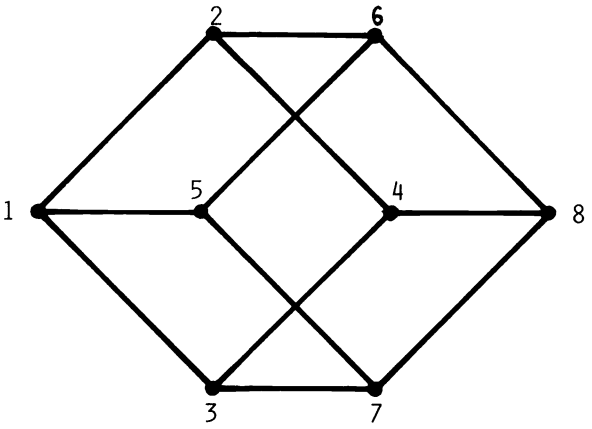


FIG. 1

Now $AXy = \alpha Le_i = \alpha Xy = XBy$. Since X has independent columns, we conclude that $By = \alpha y$. The second conclusion is proved by taking traces of both sides of (1). □

2. Symmetry of point sets.

DEFINITION. Let C be an $n \times m$ matrix of coordinates of n not necessarily distinct points in m-dimensional Euclidean space such that $C^TC = I$. The *symmetry group* of C, denoted by orth (C), is the set of $m \times m$ matrices R satisfying the condition $CR = PC$ for some permutation matrix P. This definition is based on one of Coxeter [4, p. 253].

THEOREM B. *The set* orth (C) *forms a group of orthogonal matrices. Furthermore, the set* perm (CC^T), *composed of the permutation matrices that commute with* CC^T, *is also a group and contains precisely those permutations P for which the equation* $CR = PC$ *has a solution R. Indeed, the mapping* $P \to C^TPC$ *is a group homomorphism of the second group onto the first, whose kernel is the group of permutations P that satisfy* $PC = C$.

Proof. Parts of the proof will be found in [8]; the remaining parts are routine. □

We wish to apply the ideas of the theorem above to the case where $C = Z$, the complete eigenmatrix associated with the second eigenvalue of a distance-regular graph. When interpretation in terms of the graph is desired, we speak of the vertex i corresponding to row i of Z and/or to the point in Euclidean space whose coordinates are found in that row. We need several technical lemmas.

LEMMA 2. *Let G be a connected graph, and let* α *be its second eigenvalue. Then either* $\alpha \geqq 0$ *or else G is a complete graph.*

Proof. This lemma is a minor variant on a theorem of Smith [13] cited in [5, p. 163]. □

LEMMA 3. *Let B be the coloration matrix of a distance partition of a distance-regular graph G, let* α *be its second eigenvalue, and let* $y = [y_0, y_1, \cdots, y_d]^T$ *a corresponding eigenvector with* $y_0 > 0$. *Then* $y_0 > y_1 > y_i$, $i = 2, \cdots, d$.

Proof. If G is a complete graph, then $d = 1$ and the result is trivial; from here on, assume that d is at least 2. Since B is tridiagonal, it is easy to show that y_0 cannot be zero and (from the first row of the equation $By = \alpha y$) that $y_1 = (\alpha/\rho)y_0 < y_0$. Now consider row i of the equation $By = \alpha y$. Recall that the sum of the nonzero entries in any row is ρ, and use this fact to replace the diagonal entry. Then row i reads

$$c_iy_{i-1} + (\rho - c_i - b_i)y_i + b_iy_{i+1} = \alpha y_i,$$

which is algebraically equivalent to

$$(\rho - \alpha)y_i + c_i(y_{i-1} - y_i) = b_i(y_i - y_{i+1}). \tag{2}$$

We know that y_0 is greater than y_1, which is nonnegative. Then, using $i = 1$ in (2), we see that $y_1 > y_2$. In general, if y_i is nonnegative and less than y_{i-1}, then y_{i+1} is less than y_i. Thus, the y's decrease until they reach a negative minimum. An argument developed in [6] and continued in [12] shows that the sequence of y's can change sign only once. Therefore the claimed inequalities are confirmed. □

LEMMA 4. *Let G be a distance-regular graph, and let* α *be the second eigenvalue of its adjacency matrix A. If Z is a complete eigenmatrix associated with* α, *then Z has distinct rows.*

Proof. Rows i and j of Z are equal if and only if $c^TZ = 0$, where $c = e_i - e_j$. This is true if and only if $c^TZZ^T = 0$, or $c^TL = 0$—that is, if rows i and j of L are equal. However, from (1) we see that row i of L has y_0 in column i, while row j has some y_k, $k > 0$, in column i. By Lemma 3, these numbers are different. □

THEOREM 1. *Let G be a distance-regular graph, and let Z be a complete eigenmatrix of the adjacency matrix A associated with the second eigenvalue α. Then the symmetry group* orth (Z) *is isomorphic to the automorphism group of G.*

Proof. Lemma 4 shows that the rows of Z are distinct, so orth (Z) is isomorphic to perm (ZZ^T) = perm (L) by Theorem B. Now consider (1). According to Lemma 3, the coefficient of $A_1 = A$ is greater than any of the subsequent coefficients. Therefore, a permutation matrix that commutes with L must also commute with A. On the other hand, L is a polynomial in A, so any matrix that commutes with A must also commute with L. Thus, perm (L) = perm (A), and the latter is well known to be isomorphic to the automorphism group of G. □

This theorem sharpens results of Babai [1] and Godsil [8] for the class of distance-regular graphs. A convenient example is supplied by the cube in Fig. 1, whose second eigenvalue is 1, with multiplicity 3. A complete eigenmatrix associated with 1, shown below, is made up of the coordinates of the vertices of a cube. Frucht [7] showed that the group of the skeleton of the cube is precisely what we have called orth (Z), thus confirming the results of the theorem:

$$Z=\frac{1}{\sqrt{8}}\begin{bmatrix}1 & 1 & 1 & 1 & -1 & -1 & -1 & -1\\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1\\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1\end{bmatrix}^T.$$

3. Rows of an eigenmatrix. Theorem 1 reveals a significant way in which the rows of an eigenmatrix reflect properties of the distance-regular graph to which it belongs. In this section, we find other properties of the graph that are tied to properties of the rows of an eigenmatrix. In all the theorems of this section, we assume that G is distance-regular, α is its second eigenvalue, the multiplicity of α is m, Z is a complete eigenmatrix associated with α, and $L = ZZ^T$ is the associated projector.

THEOREM 2. *Let $w_i^T = e_i^T Z$ be the ith row of Z. Then, for any vertex i of G, the minimum of $\|w_i - w_j\|$, $i \neq j$, is achieved at precisely those vertices j that are adjacent to i in G.*

Proof. The square of the quantity to be minimized is

$$\|w_i\|^2 + \|w_j\|^2 - 2w_i^T w_j.$$

However, we know that $w_i^T w_j$ is the i,j-entry of $ZZ^T = L$. From (1), we see that this quantity is $2y_0 - 2y_r$, where r = dist (i, j). By Lemma 3, this is minimized when $r = 1$, which is the conclusion of the theorem. □

In [9], Godsil proves the same theorem (his Lemma 5.4) and also a lemma (his Lemma 5.3) that generalizes Lemma 4 above. The proofs used here are independent of Godsil's.

Theorem 2 tells us that proximity of the rows of Z corresponds exactly to adjacency in G. The next theorem connects linear independence of rows to mutual adjacency.

THEOREM 3. *Let $G \neq K_n$, and let U be a set of q mutually adjacent vertices of G. Then the corresponding rows of Z are linearly independent matrices.*

Proof. First note that the set of rows in question, namely w_i^T for i in U, contains q distinct elements, by Lemma 4. Moreover, they are independent if and only if the matrices Le_i, i in U, are independent. Now, the column matrix Le_i has, by (1), a y_0 in entry i and a y_1 in entry j, if i and j are adjacent. Thus, the submatrix of L whose row and column indices are in U has y_0 on the diagonal and y_1 elsewhere, forming the $q \times q$ matrix (e is a column of ones):

$$y_0 I + y_1(ee^T - I) = (y_0 - y_1)I + y_1 ee^T$$

with eigenvalues $y_0 - y_1$ (multiplicity $q - 1$) and $y_0 + (q - 1)y_1$ (multiplicity 1). The former is nonzero by Lemma 3. The latter can be zero only if y_1 is negative, which is impossible by Lemma 2, since $G \neq K_n$. □

COROLLARY. *Let $G \neq K_n$ be distance-regular. If G contains a q-clique, then the multiplicity of the second eigenvalue is at least q.*

This corollary improves, for the second eigenvalue, the lower bound given in [14] for the multiplicity of any eigenvalue (other than ρ) of a distance-regular graph of girth 3. While Theorem 3 identifies some independent rows of Z, Theorem 4 will identify some dependent rows.

THEOREM 4. *If i is any vertex of G, and Z is a complete eigenmatrix associated with any eigenvalue α of A, then the rows of Z corresponding to i and its neighbors are dependent.*

Proof. Row i of the adjacency matrix is $e_i^T A = \sum_{j@i} e_j^T$ (where @ means "adjacent to"). Thus row i of the equation $AZ = \alpha Z$ reads

$$\sum_{j@i} w_j^T = \alpha w_i^T,$$

which certainly implies dependence. □

4. Convex polytopes. Godsil [8] suggested studying the m-polytope $P(\alpha)$ that is the convex hull of the points whose coordinates are the rows of a complete $n \times m$ eigenmatrix Z associated with an eigenvalue α of A. In the case of a distance-regular graph, all the rows w_i^T have the same norm, so all the distinct rows will be extreme points of their convex hull. (We use the term "extreme point" instead of the more usual "vertex," reserving the latter for graph usage.) Our reference for convex polytopes is [3].

A facet F of such a polytope P is the intersection of P with a hyperplane

$$H(u;\gamma) = \{x^T : x^T u = \gamma\} \qquad (\gamma \geqq 0)$$

subject to the following conditions: (1) all points x of P satisfy $x^T u \leqq \gamma$; (2) $x^T u = \gamma$ for at least m extreme points; (3) the affine dimension of F is $m - 1$. The extreme points of the facet are the extreme points of P for which equality holds. For any eigenvalue $\alpha \neq \rho$, we have $e^T Z = 0$; hence 0 is an interior point, and γ is positive.

In terms of the eigenmatrix Z, a hyperplane $H(u;\gamma)$ defining a facet can be identified by the condition: $Zu \leqq \gamma e$ with equality for at least m independent rows of Z. In this case we call $f = Zu$ a *facet vector* of P. Obviously, a facet vector of $P(\alpha)$ is also an eigenvector of A associated with α and is independent of the choice of the eigenmatrix Z.

THEOREM 5. *Let $G \neq K_n$ be a distance-regular graph, and let Z be a complete eigenmatrix associated with α, the second eigenvalue of A. If G contains a q-clique, and q is the multiplicity of α, then the rows of Z corresponding to a q-clique are extreme points of a facet of $P(\alpha)$, and that facet is a simplex.*

Proof. Lemma 4 guarantees that the rows of Z are distinct. Let U be a set of vertices that form a q-clique, and let f be the sum of the columns of $L = ZZ^T$ whose indices are in U. Then entry i of f is $f_i = y_0 + (q - 1)y_1 = \gamma$ if i is in U, but $f_i \leqq qy_1 < \gamma$ if i is not in U. Clearly γ is positive (see Lemma 2 and the proof of Lemma 3), and the rows of Z corresponding to vertices of U are independent by Theorem 3; thus f is a facet vector. Now $P(\alpha)$ is q-dimensional, and its facets are $(q - 1)$-dimensional. Since the facets we have constructed have q extreme points, they must be simplices. □

THEOREM 6. *Let G be a distance-regular graph, and let Z be a complete eigenmatrix associated with α, the second eigenvalue of A. Let U be a set composed of a vertex i and*

all its neighbors. Then the rows of Z corresponding to U are not the extreme points of a facet of the polytope $P(\alpha)$.

Proof. Let r^T be row i of $A - \alpha I$: r has $-\alpha$ in entry i and ones in the entries corresponding to neighbors of i. Then $r^T Z = 0$, and consequently $r^T f = 0$ for any facet vector f. But if $f_j = 1$ for all j in U, then $r^T f = \rho - \alpha \neq 0$, so no such facet vector can exist. □

COROLLARY. *Let U be a set composed of all the neighbors of a vertex i. Then the rows of Z corresponding to U are not the extreme points of a facet of $P(\alpha)$.*

Proof. With r as above, suppose f is a column vector such that $f_j = 1$ for all j in U. Then $r^T f \neq 0$, so f cannot be a facet vector. □

5. Antipodal graphs.

DEFINITION. A distance-regular graph is *antipodal* if, for each vertex i, there is a unique vertex i' whose distance from i is the diameter d of the graph.

This definition differs from that of [2, p. 151]. By way of an example, note that all the platonic solids except the tetrahedron have antipodal skeletons.

LEMMA 5. *Let $V_0, V_1, \cdots, V_d$ be the distance partition starting from vertex i of a distance-regular graph G. Then V_d is a singleton $\{i'\}$ if and only if the distance partition starting from i' is given by $V'_k = V_{d-k}$, for $k = 0, 1, \cdots, d$.*

Proof. Suppose the distance partition starting from some i' is given by $V'_k = V_{d-k}$, $k = 0, 1, \cdots, d$. Then $\{i'\} = V'_0 = V_d$.

Suppose next that $V_d = \{i'\}$, and let j be a vertex in V_k, so that dist $(i, j) = k$. By distance-regularity, j is adjacent to at least one vertex j_{k+1} in V_{k+1}, which is adjacent to at least one vertex j_{k+2} in V_{k+2}, etc. Thus we may construct a path from j to i', the sole vertex in V_d, having length $d - k$. This path must be minimal, for otherwise the triangle inequality is violated. Thus V_k is contained in V'_{d-k}. The reverse containment follows by a symmetric argument. □

LEMMA 6. *Let i be a vertex in an antipodal distance-regular graph G, and let* dist $(i, i') = d$, *the diameter of G. If j is any vertex of G, then* dist (i, j) + dist $(j, i') = d$.

Proof. The proof follows from the proof of Lemma 5. □

THEOREM 7. *Let G be distance-regular with distance coloration matrix B. Then B is centrosymmetric if and only if G is antipodal.*

Proof. A centrosymmetric matrix is one that commutes with the permutation matrix S that has ones on the secondary diagonal; in terms of the elements of B, centrosymmetry means that $c_i = b_{d-i}$ and $a_i = a_{d-i}$ for $i = 0, 1, \cdots, d$. Biggs [2, p. 140] gives the formula

$$k_i = b_0 b_1 \cdots b_{i-1} / c_1 c_2 \cdots c_i$$

for the number of vertices in set V_i of a distance partition. It is easy to prove that centrosymmetry implies that $k_i = k_{d-i}$, and in particular, that $k_d = 1$.

If G is antipodal, let i and i' be at distance d, X_i and let $X_{i'}$ be the indicators of the distance partitions starting at i and i'. Then the identity $X_i S = X_{i'}$ follows immediately from Lemma 5. From the coloration equation $AX_i = X_i B$ we have

$$\begin{aligned} AX_i S &= X_i BS \\ &= AX_{i'} = X_{i'} B = X_i SB. \end{aligned}$$

Then, because the columns of X_i are independent, $BS = SB$ follows. □

LEMMA 7. *Let G be distance-regular and antipodal. Then $A_d A_k = A_{d-k}$, for $k = 0, 1, \cdots, d$.*

Proof. The distance matrix A_d has a 1 in the i,j-position if and only if dist $(i, j) = d$. Thus, A_d is a permutation matrix that interchanges each vertex i with its antipode. Now, row i of the product A_dA_k is row i' of A_k, so it has a 1 in entry j if and only if dist $(j, i') = k$, which means that dist $(i, j) = d - k$, by Lemma 6. □

THEOREM 8. *Let G be distance-regular and antipodal. Then* (i) *the interchange of each vertex with its antipode is an automorphism of G*; (ii) *the polytope associated with the second eigenvalue of G admits central inversion as a symmetry*; (iii) *the facets of this polytope occur in parallel pairs.*

Proof. (i) As observed in the proof of Lemma 7, the distance matrix A_d is a permutation matrix that interchanges each vertex with its antipode. Since A_d is a polynomial in A, it commutes with A and thus represents an automorphism of G.

(ii) Now let $P = A_d$. We need to show that $PZ = -Z$, or equivalently that $PL = -L$, where $L = ZZ^T$. By (1) and Lemma 7,

$$A_dL = y_0A_d + y_1A_{d-1} + \cdots + y_dA_0.$$

Now, the eigenvector y of B is unique when normalized by $y_0 = m/n$. Because S commutes with B, any eigenvector of B is also an eigenvector of S: $Sy = \pm y$. From Lemma 3, we know that $y_d < 0$, so $Sy = -y$, or $y_k = -y_{d-k}$ for $k = 0, 1, \cdots, d$. Thus $PL = -L$, as required.

(iii) If f is a facet vector, then $Pf = -f$ is one also. □

6. Examples. In this section we list some distance-regular graphs and families of graphs for which the polytopes associated with the second eigenvalue can be determined. For each graph, the following information is needed: the coloration matrix B, its eigenvectors in the form of a square matrix Y, the spectrum of A in the style of [2] with eigenvalues above their multiplicities, and the number of vertices in each set of the distance partition as a column k.

(1) $G = K_n$, the complete graph. The essential information is shown below:

$$B = \begin{bmatrix} 0 & n-1 \\ 1 & n-2 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 & n-1 \\ 1 & -1 \end{bmatrix}, \quad k = \begin{bmatrix} 1 \\ n-1 \end{bmatrix},$$

$$\text{spec } (A) = \begin{Bmatrix} n-1 & -1 \\ 1 & n-1 \end{Bmatrix}.$$

The eigenvector information shows that each column of the projector L has (aside from a normalizing factor of $1/n$) an $n - 1$ in the diagonal position and a -1 in all other positions; that is, $L = (nI - ee^T)/n$. A complete eigenmatrix Z is composed of $n - 1$ mutually orthogonal columns, each orthogonal to e. Since all off-diagonal elements of $L = ZZ^T$ are the same, the n rows of Z are equidistant. Thus the polytope $P(-1)$ is a simplex.

(2) $G = K_{2m} - mK_2$, $m \geqq 2$. Suppose that edges $\{1, m + 1\}, \cdots, \{m, 2m\}$ are deleted from the complete graph on $2m$ vertices. The resulting graph is distance-regular with diameter 2 and is described by the following information:

$$B = \begin{bmatrix} 0 & 2(m-1) & 0 \\ 1 & 2(m-2) & 1 \\ 0 & 2(m-1) & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 1 & 1 & m-1 \\ 1 & 0 & -1 \\ 1 & -1 & m-1 \end{bmatrix}, \quad k = \begin{bmatrix} 1 \\ 2(m-1) \\ 1 \end{bmatrix},$$

$$\text{spec } (A) = \begin{Bmatrix} 2m-2 & 0 & -2 \\ 1 & m & m-1 \end{Bmatrix}.$$

In G, each vertex is a member of 2^{m-1} different m-cliques and no larger clique. By Theorem 6, the vertices of each clique correspond to the extreme points of a facet of $P(0)$. From the eigenvector information, we can see that a complete eigenmatrix associated with the eigenvalue 0 is

$$Z=\frac{1}{\sqrt{2}}\begin{bmatrix} I \\ -I \end{bmatrix}.$$

Since the rows of Z are the coordinates of the m-dimensional cross-polytope (see [4, p. 122]), which has exactly the facets described, we conclude that the description of $P(0)$ is complete.

(3) $G = K_{m,m}$, the complete bipartite graph. We require that $m \geqq 2$, so that the diameter is 2. The essential information is

$$B=\begin{bmatrix} 0 & m & 0 \\ 1 & 0 & m-1 \\ 0 & m & 0 \end{bmatrix}, \quad Y=\begin{bmatrix} 1 & m-1 & 1 \\ 1 & 0 & -1 \\ 1 & -1 & 1 \end{bmatrix}, \quad k=\begin{bmatrix} 1 \\ m \\ m-1 \end{bmatrix},$$

$$\text{spec}\,(A)=\begin{Bmatrix} m & 0 & -m \\ 1 & 2m-2 & 1 \end{Bmatrix}.$$

Assume that the two parts contain the vertices $1, 2, \cdots, m$ and $m+1, m+2, \cdots, 2m$, respectively. Then the adjacency matrix A and the projector associated with zero are

$$A=\begin{bmatrix} 0 & ee^T \\ ee^T & 0 \end{bmatrix}, \quad L=\frac{1}{m}\begin{bmatrix} mI-ee^T & 0 \\ 0 & mI-ee^T \end{bmatrix}.$$

Each column of mL contains zeros in m entries, $m-1$ in one entry, and -1 in the remaining $m-1$ entries. Thus, the product of mL by the column

$$g_{ij}=-\begin{bmatrix} e_i \\ e_j \end{bmatrix}$$

is an eigenvector with $-2(m-1)$ in two entries and 1 in the remaining $2m-2$ entries. Since the multiplicity of zero is $2m-2$, this product is a facet vector, and the facet is a simplex, which is "opposite" a pair of adjacent vertices, i and $m+j$. Clearly there are m^2 such facets.

Any facet must contain at least $2m-2$ vertices. However, by Theorem 5, no facet can contain a vertex and all its neighbors. Therefore, we have found all the facets of $P(0)$, which is consequently simplicial.

(4) $G_m = L(K_m)$, the line graph of K_m. This graph is also known as a triangular graph because the number of vertices is the triangular number $t_m = m(m-1)/2$. We require that $m \geqq 4$, so that the diameter will be 2:

$$B=\begin{bmatrix} 0 & 2m-4 & 0 \\ 1 & m-2 & m-3 \\ 0 & 4 & 2m-8 \end{bmatrix}, \quad Y=\begin{bmatrix} 1 & 2m-4 & (m-2)(m-3) \\ 1 & m-4 & -(m-3) \\ 1 & -4 & 2 \end{bmatrix},$$

$$\text{spec}\,(A)=\begin{Bmatrix} 2m-4 & m-4 & -2 \\ 1 & m-1 & m(m-3)/2 \end{Bmatrix}, \quad k=\begin{bmatrix} 1 \\ 2m-4 \\ (m-2)(m-3)/2 \end{bmatrix}.$$

It is helpful to think of the vertices as unordered pairs, $\{a, b\}$, $a \neq b$, $a, b = 1, \cdots, m$. Then $\{a, b\}$ is adjacent to $\{a, x\}$ for any $x \neq b$. Indeed, we see that the set of all

vertices of the form $\{a, x\}$, $x \neq a$, form an $(m-1)$-clique. By Theorem 5, these are extreme points of a facet that is a simplex, and there are m of these. Closer inspection shows the facet vector to have $m-2$ in the $m-1$ entries corresponding to the vertices of a clique and -2 in the remaining $(m-1)(m-2)/2$ entries. Since $m \geqq 4$, the negative of this vector is the facet vector of a different facet, supplying m more facets.

To find further facets of $P(m-4)$, note that the negative of the column of L that corresponds to a vertex $\{a, b\}$ has (see Y above) the same positive number in the $(m-2)(m-3)/2$ places corresponding to vertices at distance 2 from $\{a, b\}$ and lesser numbers in the rest. If $(m-2)(m-3)/2 \geqq m-1$, that is if $m \geqq 6$, such a column is a positive multiple of a facet vector. This construction produces $n = m(m-1)/2$ new facets.

There are two interesting special cases. First, G_4 is the skeleton of the octahedron, which is also the special case $m = 3$ of item 2. We have found the eight facets, in four parallel pairs. Second, G_5 is the complement of Petersen's graph. The facets, which are three-dimensional, are five tetrahedra paired with five octahedra. By direct computation, it has been shown that there are no more facets. For G_m, $m \geqq 6$, it is not yet known whether we have found all the facets.

(5) $G = Q_d$, the d-dimensional cube. The graph is distance-regular with diameter d. The spectrum consists of the numbers $d-2i$, with multiplicity $d!/i!(d-i)!$, for $i = 0, 1, \cdots, d$ (see [2, pp. 138, 145]). In [11] it is shown by induction that the polytope $P(d-2)$ is the d-dimensional cube itself. The information about the coloration matrix of G given in [2, p. 138] shows these graphs to be antipodal, and, of course, they admit central inversion as a symmetry.

Acknowledgment. The author thanks the reviewer for a most careful reading and for suggesting several improvements in the proofs.

REFERENCES

[1] L. BABAI, *Automorphism group and category of cospectral graphs*, Acta Math. Acad. Sci. Hungar., 31 (1978), pp. 295–306.
[2] N. BIGGS, *Algebraic Graph Theory*, Cambridge Univ. Press, London, 1974.
[3] A. BRONDSTED, *An Introduction to Convex Polytopes*, Springer-Verlag, New York, Berlin, 1983.
[4] H. S. M. COXETER, *Regular Polytopes*, 3rd edition, Dover, New York, 1973.
[5] D. M. CVETKOVIĆ, M. DOOB, AND H. SACHS, *Spectra of Graphs*, VEB, Berlin; Academic Press, New York, 1980.
[6] M. FIEDLER, *A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory*, Czechoslovak Math. J., 25 (1975), pp. 619–633.
[7] R. FRUCHT, *Die Gruppe des Petersen'schen Graphen und der Kantensysteme der regulaeren Polyeder*, Comment. Math. Helv., 9 (1936), pp. 217–223.
[8] C. GODSIL, *Graphs, groups and polytopes*, in *Combinatorial Mathematics VI (Canberra 1977)*, D. A. Holton and J. Seberry, eds., Springer-Verlag, Berlin, New York, 1978, pp. 157–164.
[9] ———, *Bounding the diameter of distance-regular graphs*, Combinatorica, to appear.
[10] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, 2nd edition, Academic Press, New York, 1985.
[11] C. LICATA AND D. L. POWERS, *A surprising property of some regular polytopes*, Scientia, to appear.
[12] D. L. POWERS, *Structure of a matrix according to its second eigenvector*, in Current Trends in Matrix Theory, F. Uhlig and R. Grone, eds., North-Holland, New York, Amsterdam, 1987, pp. 261–266.
[13] J. H. SMITH, *Some properties of the spectrum of a graph*, in Combinatorial Structures and Their Applications, R. Guy et al., eds., Gordon and Breach, New York, 1970, pp. 403–406.
[14] P. TERWILLIGER, *Eigenvalue multiplicities of highly symmetric graphs*, Discrete Math., 41 (1982), pp. 295–302.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

013

INFLATION MATRICES AND *ZME*-MATRICES THAT COMMUTE WITH A PERMUTATION MATRIX*

JEFFREY L. STUART†

Abstract. Centrosymmetric matrices are matrices that commute with the permutation matrix J, the matrix with ones on its cross-diagonal. This paper generalizes the concept of centrosymmetry, and considers the properties of matrices that commute with an arbitrary permutation matrix P, the P-commutative matrices. In particular, it focuses on two related classes of matrices: inflation matrices and *ZME*-matrices. The structure of P-commutative inflators is determined, and then this is used to characterize the P-commutative *ZME*-matrices. Centrosymmetric matrices in these classes are presented as a special case.

Key words. centrosymmetric matrix, P-commutative matrix, inflation matrix, *ZME*-matrix

AMS(MOS) subject classifications. primary 15A27; secondary 15A48, 15A57

1. Introduction. The concept of a matrix that commutes with a fixed permutation matrix is a natural generalization of the concept of a *centrosymmetric matrix*, a matrix which commutes with the $n \times n$ permutation matrix J given by $J = [\delta_{i,n-i+1}]$, where δ_{ij} is the Kronecker delta. The square matrix A is called a *P-commutative matrix* if $AP = PA$, or equivalently if $P^{-1}AP = A$. If $p(i)$ denotes the permutation map on $\{1, 2, \cdots, n\}$ corresponding to the $n \times n$ permutation matrix P, then A is P-commutative if and only if $A_{ij} = A_{p(i)p(j)}$ for $1 \leqq i, j \leqq n$. In particular, by $P = J$, the $n \times n$ matrix A is centrosymmetric if and only if $A_{ij} = A_{n-i+1,n-j+1}$ for all i and j.

Centrosymmetric matrices have arisen in the study of symmetric Toeplitz matrices [3], and in applications of Markov processes to genetics [4]. The basic properties of centrosymmetric matrices are summarized in [6].

The literature devoted to Z-matrices and M-matrices is quite extensive. In a recent paper [1], Friedland, Hershkowitz, and Schneider study a certain class of irreducible Z-matrices each of whose positive integer powers is again a Z-matrix subject to certain irreducibility conditions. The authors characterize this class, called the *ZME*-matrices, in terms of sequences of certain matrices called inflators and a new matrix product called inflation.

This paper studies the structure of the P-commutative members of the class of *ZME*-matrices, where P is a permutation matrix, by determining the behavior of P-commutative matrices under the inflation product and by determining the structure of P-commutative inflators. The relation between P-commutativity and inflation is determined in §§ 3 and 4, and the main results are Theorems 4.1 and 4.11. The structure of P-commutative *ZME*-matrices is developed in §§ 5 and 6, and the principal result is Theorem 6.1.

2. *P*-commutative matrices. Let $\mathcal{M}_n(\mathbb{C})$ denote the set of $n \times n$ matrices over $\mathbb{C}$. If P is in $\mathcal{M}_n(\mathbb{C})$, let spec (P) denote the set of distinct eigenvalues of P. Let Eig $(P) = \{x \in \mathbb{C}^n: Px = \lambda x \text{ for some } \lambda \text{ in spec } (P)\}$. The elements of Eig (P) are called *P-commutative vectors*.

The following lemma summarizes well-known properties of permutation matrices.

* Received by the editors December 8, 1986; accepted for publication (in revised form) December 21, 1987.

† Department of Mathematics, University of Southern Mississippi, Hattiesburg, Mississippi 39406.

LEMMA 2.1. *Let P be an $n \times n$ permutation matrix. Then the eigenvalues of P are nth roots of unity. Further, P has n distinct eigenvalues if and only if P is irreducible, or equivalently, if and only if the permutation corresponding to P is an n-cycle.*

THEOREM 2.2. *Let A be in $\mathcal{M}_n(\mathbb{C})$. Let P be an $n \times n$ permutation matrix.*

(i) *If A is P-commutative, then every eigenspace for A has a basis of vectors in* Eig (P).

(ii) *If A has n linearly independent eigenvectors which are in* Eig (P), *then A is P-commutative.*

(iii) *If P is irreducible, and if A is P-commutative, then A is diagonalizable.*

Proof. These are standard results from matrix theory. The proofs of (ii) and (iii) can be found in [5, pp. 264–265]. To prove (i), note that $AP = PA$ implies that P maps an eigenspace of A into itself. Let V be an eigenspace of A. Since P is diagonalizable and nonsingular, the linear transformation which is P restricted to V has a full set of linearly independent eigenvectors which form a basis for V. □

LEMMA 2.3. *Let P be an $n \times n$ permutation matrix. The set of $n \times n$ matrices which commute with A form an algebra over $\mathbb{C}$.*

Proof. This is an elementary result which can be directly verified. □

The following result will be used repeatedly in subsequent sections.

THEOREM 2.4. *Let A be in $\mathcal{M}_m(\mathbb{C})$. Let P be an $n \times n$ permutation matrix. Suppose that A is a P-commutative matrix. Suppose that A has k distinct eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_k$. Then there exist k complex, P-commutative, idempotent $n \times n$ matrices E_i and there exist k complex, P-commutative, nilpotent $n \times n$ matrices Z_i such that*

$$A = \sum_{i=1}^{k} (\lambda_i E_i + Z_i) \tag{2.5}$$

and $E_iE_j = E_iZ_j = Z_iE_j = Z_iZ_j = 0$ when $i \neq j$. Finally, if the matrix A is diagonalizable, then each Z_i is the zero matrix.

Proof. From the standard theory of spectral decomposition of matrices (see [2, p. 100 ff.]) it is well known that every complex $n \times n$ matrix A with k distinct eigenvalues has a decomposition, as in (2.5), where the E_i are idempotent and the Z_i are nilpotent, where the E_i and Z_i satisfy the various product relationships listed in the statement of the theorem, and where $E_i = r_i(A)$ and $Z_i = s_i(A)$, where $r_i(x)$ and $s_i(x)$ are certain polynomials which divide the minimal polynomial of A. Since A is a P-commutative matrix, and since the P-commutative matrices form an algebra, it follows that each E_i and each Z_i is a P-commutative matrix. Finally, it is known that if A is diagonalizable, then $s_i(A) = 0$ for each i. □

COROLLARY 2.6. *Let A be in $\mathcal{M}_n(\mathbb{R})$. Suppose that A has real spectrum. Then the matrices E_i and Z_i given in the preceding theorem are real.*

Proof. Since the spectrum of A is real, every factor of the minimum polynomial of A is real; consequently the polynomials $r_i(x)$ and $s_i(x)$ have real coefficients. Since A is real, the result is clear. □

A *strictly nonzero matrix* (*strictly nonzero vector*) is a matrix (vector) each of whose entries is nonzero. A *strictly positive matrix* (*strictly positive vector*) is a matrix (vector) each of whose entries is positive.

LEMMA 2.7. *Let U be a strictly nonzero $n \times n$ matrix. Let P be an $n \times n$ permutation matrix. Then the following are equivalent*:

(i) *U is a rank one, P-commutative matrix.*

(ii) *There exist strictly nonzero vectors u and v such that $U = uv^t$ and such that u and v are eigenvectors for P corresponding to reciprocal eigenvalues.*

(iii) *U is rank one, and for every pair of strictly nonzero vectors u and v such that $U = uv^t$, it follows that u and v are eigenvectors for P corresponding to reciprocal eigenvalues.*

Proof. Lemma 2.7 (iii) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (i) are obvious. It remains to show that (i) implies (iii). Suppose that u and v are strictly nonzero vectors such that $U = uv^t$. Since $P^tUP = U$, it follows that for all α and β, $u_\alpha v_\beta = u_{p(\alpha)}v_{p(\beta)}$. Fixing β, it is clear from the strict nonzero condition on v that $u_\alpha[u_{p(\alpha)}]^{-1}$ is a nonzero constant independent of α. Call this constant γ. Then $Pu = \gamma^{-1}u$. Then the strict nonzero condition on u implies $Pv = \gamma v$. That is, u and v are eigenvectors for P corresponding to reciprocal eigenvalues. □

The following corollary is immediate.

COROLLARY 2.8. *The matrix U in the preceding lemma is strictly positive if and only if u and v can be chosen to be strictly positive eigenvectors of P corresponding to the eigenvalue one.*

3. Inflation. In this section, the concept of inflation introduced in [1] is discussed.

Let m and n be positive integers with $m \leqq n$. An *m-partition of n* is a partition of the set $\{1, 2, \cdots, n\}$ into an ordered collection of m nonempty, disjoint sets such that the elements within each set are arranged in ascending order. Let Π be the m-partition of n given by the ordered collection $B_1, B_2, \cdots, B_m$. Let Q be an $m \times m$ permutation matrix. Let P be an $n \times n$ permutation matrix. Let $q(i)$ and $p(i)$ be the permutations corresponding to Q and P, respectively. The partition Π is called a *Q, P-commutative partition* if for each r with $1 \leqq r \leqq m$ and for each i in B_r, the index $p(i)$ is in $B_{q(r)}$. Finally, the partition Π is called a *centrosymmetric partition* if $P = J_n$.

Let m and n be positive integers with $m \leqq n$. Let Π be an m-partition of n given by $B_1, B_2, \cdots, B_m$. Let v be a vector in $\mathbb{C}^n$. Then Π partitions v into m blocks. Let $v_{\langle j \rangle}$ denote the subvector of v consisting of the entries of v which are indexed by B_j. Let U be an $n \times n$ matrix. Then Π induces a block-partitioning of U. Let $U_{\langle i,j \rangle}$ denote the block of U consisting of the entries of U whose indices are in $B_i \times B_j$.

Let m and n be positive integers with $m \leqq n$. Let A be an $m \times m$ matrix. Let U be an $n \times n$ matrix. Let Π be an m-partition of n. Give U the block-partitioning induced by Π. The *inflation matrix of A by U with respect to* Π is the $n \times n$ matrix denoted by $A \times\times U$ which is defined as follows [1, Def. 4.1]: For each α and β in $\{1, 2, \cdots, n\}$, there exist unique indices r and s such that $\alpha \in B_r$ and $\beta \in B_s$; let $(A \times\times U)_{\alpha\beta} = a_{rs}U_{\alpha\beta}$.

LEMMA 3.1. *Let m and n be positive integers with $m \leqq n$. Let Q be an $m \times m$ permutation matrix. Let P be an $n \times n$ permutation matrix. Let A be an $m \times m$ Q-commutative matrix. Let U be an $n \times n$ P-commutative matrix. Let Π be a Q, P-commutative m-partition of n. Give U the block-partitioning induced by Π. Then $A \times\times U$ is a P-commutative matrix.*

Proof. Let Π be given as in the definition of a Q, P-commutative partition. Consider the α, β entry of $A \times\times U$. There exist unique r and s in $\{1, 2, \cdots, m\}$ such that $\alpha \in B_r$ and $\beta \in B_s$. Compute $(A \times\times U)_{\alpha\beta} = a_{rs}U_{\alpha\beta} = a_{q(r),q(s)}U_{p(\alpha),p(\beta)}$ since A and U are, respectively, Q- and P-commutative matrices. Since Π is Q, P-commutative, $p(\alpha) \in B_{q(r)}$ and $p(\beta) \in B_{q(s)}$. Thus $(A \times\times U)_{p(\alpha),p(\beta)} = a_{q(r),q(s)}U_{p(\alpha),p(\beta)} = (A \times\times U)_{\alpha\beta}$. That is, $A \times\times U$ is a P-commutative matrix. □

LEMMA 3.2. *Let m and n be positive integers with $m \leqq n$. Let A be in $\mathcal{M}_m(\mathbb{C})$. Let U be in $\mathcal{M}_n(\mathbb{C})$. Let Q be an $m \times m$ permutation matrix. Let P be an $n \times n$ permutation matrix. Suppose that U is a P-commutative matrix. Let Π be a Q, P-commutative m-partition of n. Let U be partitioned by Π. Suppose that U has no zero blocks in this partitioning. If $A \times\times U$ is P-commutative, then A is Q-commutative.*

Proof. Choose α and β in $\{1, 2, \cdots, n\}$. Let r and s be the unique indices such that $\alpha \in B_r$ and $\beta \in B_s$. Since Π is Q, P-commutative, $p(\alpha) \in B_{q(r)}$ and $p(\beta) \in B_{q(s)}$. Since U is P-commutative and has no zero blocks, $0 \neq U_{\alpha\beta} = U_{p(\alpha),p(\beta)}$. Since $A \times\times U$ is P-commutative and since Π is Q, P-commutative, $a_{rs}U_{\alpha\beta} = (A \times\times U)_{\alpha\beta} = (A \times\times U)_{p(\alpha),p(\beta)} = a_{q(r),q(s)}U_{p(\alpha),p(\beta)}$. Thus $a_{rs} = a_{q(r),q(s)}$. Since α and β are arbitrary, it follows that the final equality holds for all r and s. That is, A is a Q-commutative matrix. □

4. Inflators and $G(U)$. The following definition of a (normalized) inflator is Definition 4.3 of [1].

Let m and n be positive integers with $m \leqq n$. Let Π be an m-partition of n. Let U be a strictly nonzero matrix in $\mathcal{M}_n(\mathbb{C})$. The matrix U is called an *inflator (associated with* Π) if there exist a pair of strictly nonzero column vectors u and v in $\mathbb{C}^n$ such that the following conditions hold for the subvectors of u and v corresponding to the partition blocks of Π:

(i) For each i and j in $\{1, 2, \cdots, m\}$, $U_{\langle i,j\rangle} = u_{\langle i\rangle}[v_{\langle j\rangle}]^t$;

(ii) For each j with $1 \leqq j \leqq m$, $[v_{\langle j\rangle}]^t[u_{\langle j\rangle}] = 1$. (Note that condition (i) implies that $U = uv^t$.) An inflator U is called a *normalized inflator* if u and v can be chosen so that they satisfy a third condition:

(iii) For each i with $1 \leqq i \leqq m$, $[u_{\langle i\rangle}]^*[u_{\langle i\rangle}] = [v_{\langle i\rangle}]^*[v_{\langle i\rangle}]$.

If there exists an $m \times m$ permutation matrix Q and an $n \times n$ permutation matrix P such that the inflator U is P-commutative as a matrix and such that Π is a Q, P-commutative partition, then U is called a *Q, P-commutative inflator.* Finally, if U is a centrosymmetric matrix and if Π is a centrosymmetric partition, then U is called a *centrosymmetric inflator.* (Note that in this case, the existence of the permutation matrix Q is implicit.)

Let $\{U_i\}_{i=1}^k$ be a sequence of matrices such that U_1 is the 1×1 zero matrix, and such that for each i with $1 < i \leqq k$, U_i is an inflator and the number of blocks in the partition corresponding to U_i equals the order of U_{i-1}. Then $\{U_i\}_{i=1}^k$ is called an *inflation sequence.* If for each $i \geqq 2$, U_i is a strictly positive matrix, then the sequence is a *strictly positive inflation sequence.* Finally, if for $1 < i \leqq k$, the matrices are normalized inflators, then the sequence is called a *normalized inflation sequence.*

THEOREM 4.1. *Let U be an inflator associated with the m-partition Π of n, where $n \geqq 2$. Let Q be an $m \times m$ permutation matrix. Let P be an $n \times n$ permutation matrix. Then U is a Q, P-commutative inflator if and only if Π is a Q, P-commutative partition and $U = uv^t$, where u and v are strictly nonzero eigenvectors for P corresponding to reciprocal eigenvalues.*

Additionally, if U is Q, P-commutative, if p is the permutation corresponding to P, and if λ is the eigenvalue of P such that $Pu = \lambda u$, then for each i and j,

$$U_{ij} = \lambda U_{i,p(j)} = \lambda^{-1}U_{p(i),j} = U_{p(i),p(j)}.$$

Proof. If U is a Q, P-commutative inflator, then Π is necessarily a Q, P-commutative partition. So assume that Π is a Q, P-commutative partition. Since U is a rank one, strictly nonzero matrix, u and v exist with the desired properties by Lemma 2.7.

Conversely, if U is given in terms of u and v, then U is a strictly nonzero, rank one, P-commutative matrix by Lemma 2.7. Since Π is a Q, P-commutative partition, it follows that U is a Q, P-commutative inflator with respect to Π.

The relation among the entries follows immediately from $PUP^t = U$, and from $Pu = \lambda u$ and $Pv = \lambda^{-1}v$. □

COROLLARY 4.2. *Let U be a strictly positive inflator associated with the m-partition Π of n where $n \geqq 2$. Then U is a centrosymmetric inflator if and only if Π is a centro-*

symmetric partition and $U = uv^t$, *where* u *and* v *are strictly positive vectors satisfying* $Ju = u$ *and* $Jv = v$.

Further, if U *is such an inflator, then for each* i *and* j,

$$U_{ij} = U_{i,n-j+1} = U_{n-i+1,j} = U_{n-i+1,n-j+1}.$$

Proof. Apply the preceding theorem with $P = J$, and use Corollary 2.8. □

Let U be an inflator associated with an m-partition of n. Define the matrix $G(U)$ by $G(U) = I_n - (I_m \times\times U)$. If U is the 1×1 zero matrix, define $G(U) = I_1$. The matrix $G(U)$ is the fundamental building block in the construction of inflation-generated matrices. Its properties are developed in [1, § 5].

LEMMA 4.3. *Let* U *be a* Q, P*-commutative inflator for some permutation matrices* Q *and* P. *Then* $G(U)$ *is a* P*-commutative matrix.*

Proof. Since I_k is Q-commutative, it follows by Lemma 3.1 that $I_k \times\times U$ is P-commutative. Now use Lemma 2.3. □

LEMMA 4.4. *Let* $\{U_i\}_{i=1}^k$ *be an inflation sequence. Let* $\{P_i\}_{i=1}^k$ *be a sequence of permutation matrices such that the order of* P_i *equals the order of* U_i *for each* i. *Suppose that for each* $i \geqq 2$, U_i *is a* P_{i-1}, P_i*-commutative inflator. Suppose that* U_k *is* $n \times n$. *Let* $\Omega = \{G(U_i) \times\times U_{i+1} \times\times \cdots \times\times U_k : 1 \leqq i < k\} \cup \{G(U_k)\}$. *Then the elements of* Ω *are all* $n \times n$ P_k*-commutative matrices.*

Proof. This is an immediate consequence of the preceding lemma and repeated applications of Lemma 3.1. □

LEMMA 4.5. *Let* U *be an* $n \times n$ *inflator associated with* Π, *an* m*-partition of* n. *Suppose that* $G(U)$ *is a* P*-commutative matrix for some* $n \times n$ *permutation matrix* P. *Then there exists an* $m \times m$ *permutation matrix* Q *such that* Π *is a* Q, P*-commutative partition.*

Proof. Let $p(i)$ be the permutation corresponding to P. Let Π be given by $B_1, B_2, \cdots, B_m$. Since $G(U)$ commutes with P, so does $I \times\times U$ by Lemma 2.3. Observe that $(I \times\times U)_{\alpha\beta} \neq 0$ exactly when α and β are both in B_i for some i. Similarly,

$$(I \times\times U)_{p(\alpha)p(\beta)} \neq 0$$

exactly when both $p(\alpha)$ and $p(\beta)$ are in B_j for some j. Since $I \times\times U$ is P-commutative, $(I \times\times U)_{\alpha\beta} = (I \times\times U)_{p(\alpha)p(\beta)}$ for all α and β. Thus α and β are in B_i for some i if and only if $p(\alpha)$ and $p(\beta)$ are in B_j for some j. Since α and β are arbitrary elements of B_i, we conclude that for each i, p sends B_i to B_j for some j. That is, p acts on the sets in Π as a permutation. Hence there is a permutation q of $\{1, 2, \cdots, m\}$ such that $B_{q(r)} = \{p(\alpha) : \alpha \in B_r\}$ for $1 \leqq r \leqq m$. □

At this stage, it would be useful to prove the converse to Lemma 4.3: If $G(U)$ is P-commutative, then there exists a Q such that U is a Q, P-commutative inflator. This assertion, however, is false. Indeed, as the next example will demonstrate, not even the following, weaker assertion holds: If $G(U)$ is P-commutative, then there exist a Q and a Q, P-commutative inflator $\hat{U}$ such that $G(\hat{U}) = G(U)$.

Example 4.6. Let H be the permutation matrix

$$H = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Let $\omega = \exp(i\pi/3)$. Let x be a normalized eigenvector for H^t corresponding to ω. Let y be a normalized eigenvector for H^t corresponding to $\omega^{-1} = \bar{\omega}$. Let u and v be the partitioned vectors $u = (x^t \quad y^t)^t$ and $v = \bar{u}$. Let $U = uv^t$. Then U is a normalized inflator. Observe

that $G(U) = [xx^*] \oplus [yy^*]$. Let P be the permutation matrix $P = H \oplus H$. Then

$$P^t G(U)P = H^t[xx^*]H \oplus H^t[yy^*]H = \omega\bar{\omega}[xx^*] \oplus \bar{\omega}\omega[yy^*] = G(U).$$

Suppose that $\hat{U}$ is an inflator such that $G(\hat{U}) = G(U)$. Then since u and v are strictly nonzero, $\hat{U} = \hat{u}\hat{v}^t$ where $\hat{u} = (\alpha x^t \quad \beta y^t)^t$ for some nonzero scalars α and β, and where $\hat{v} = (\alpha^{-1}x^t \quad \beta^{-1}y^t)^t$. It is now shown that there are no choices for α and β such that $\hat{U}$ is P-commutative. Observe that if such a choice exists, then $[P^t\hat{U}P]_{\langle 1,2\rangle} = \hat{U}_{\langle 1,2\rangle}$. That is, $H^t[\alpha x\beta^{-1}y^*]H = \alpha x\beta^{-1}y^*$. Since xy^* is strictly nonzero, this last equation implies $\omega^2 = 1$, a contradiction.

The problem in the preceding example is that the subvectors of u in the preceding problem correspond to different eigenvalues of P. More specifically, it will be shown that if the subvectors of u determined by the cycles of q (the block permutation induced by P) correspond to distinct eigenvalues of P, then $\hat{U}$ cannot be found. Thus it is necessary to study the cycle structure of p.

Recall from the proof of Lemma 4.5 that if $G(U)$ is P-commutative, then the permutation p on $\{1, 2, \cdots, n\}$ corresponding to P induces a permutation q on $\{1, 2, \cdots, m\}$. In particular, the nonzero blocks of P were precisely $P_{\langle i,q(i)\rangle}$. It is the cycle structure of q that is crucial. Suppose that q consists of k disjoint cycles. Label the cycles with the numbers 1 through k. Then P has a block-partitioning induced by q. Let $P_{[i]}$ denote the submatrix of P containing all blocks $P_{\langle r,s\rangle}$ such that both r and s are in cycle i of q. Note that P is a permutation similar to the direct sum of $P_{[1]}$ through $P_{[k]}$. Let $U_{[i,j]}$ denote the submatrix of U containing all blocks $U_{\langle r,s\rangle}$ such that r is in cycle i of q and s is in cycle j. Thus $(P^tUP)_{[i,j]} = P_{[i]}^t U_{[i,j]} P_{[j]}$ for $1 \leqq i, j \leqq k$. If u is a vector in $\mathbb{C}^n$, let $u_{[i]}$ be subvector of u consisting of all the blocks $u_{\langle r\rangle}$ such that r is in cycle i of q. Thus $[Pu]_{[i]} = P_{[i]}u_{[i]}$ for $1 \leqq i \leqq k$. Suppose that U satisfies $U = uv^t$ where u and v are strictly nonzero. Then by Lemma 2.7, U is P-commutative if and only if there exists a unique λ of modulus one such that

$$P_{[i]}u_{[i]} = \lambda u_{[i]} \quad \text{and} \quad P_{[i]}v_{[i]} = \lambda^{-1}v_{[i]}$$

for $1 \leqq i \leqq k$.

LEMMA 4.7. *Let $U = uv^t$ be an inflator associated with an m-partition Π of n. Let P be an $n \times n$ permutation matrix. Suppose that $G(U)$ is P-commutative. Let Q be an $m \times m$ permutation matrix such that Π is a Q, P-partition, and let q be the permutation corresponding to Q. Then there exist nonzero $\lambda_1, \lambda_2, \cdots, \lambda_m$ such that*

$$[P^t]_{\langle q(i),i\rangle}u_{\langle i\rangle} = \lambda_i u_{\langle q(i)\rangle} \tag{4.8}$$

and

$$[v_{\langle i\rangle}]^t P_{\langle i,q(i)\rangle} = (\lambda_i)^{-1}[v_{\langle q(i)\rangle}]^t \tag{4.9}$$

for $1 \leqq i \leqq m$. Further, if q is decomposed into disjoint cycles, then

$$|\Pi_r \lambda_r| = 1$$

where r runs through the indices of any cycle in the decomposition.

Proof. Observe that $P^t[G(U)]P = G(U)$ if and only if $P^t[I_m \times\times U]P = I_m \times\times U$, that is, if and only if

$$[P^t]_{\langle q(i),i\rangle}U_{\langle i,i\rangle}P_{\langle i,q(i)\rangle} = U_{\langle q(i),q(i)\rangle}$$

for each i. In terms of the vectors u and v,

$$[P^t]_{\langle q(i),i\rangle}u_{\langle i\rangle}[v_{\langle i\rangle}]^t P_{\langle i,q(i)\rangle} = u_{\langle q(i)\rangle}[v_{\langle q(i)\rangle}]^t \tag{4.10}$$

for $1 \leqq i \leqq m$. Since u and v are strictly nonzero, this can only happen if and only if there exist nonzero numbers $\lambda_1, \lambda_2, \cdots, \lambda_m$ such that (4.8) and (4.9) both hold for all i.

Choose a cycle in the decomposition of q. Without loss of generality, the cycle is $(1, 2, 3, \cdots, h)$. Thus $q(i) = i + 1 \pmod{h}$. Then by repeated application of (4.8),

$$[P^t]_{\langle 1,h\rangle} \cdots [P^t]_{\langle 3,2\rangle}[P^t]_{\langle 2,1\rangle} u_{\langle 1\rangle} = \lambda_h \cdots \lambda_2 \lambda_1 u_{\langle 1\rangle}.$$

The left-hand side of this equation is precisely $[[P^t]^h]_{\langle 1,1\rangle} u_{\langle 1\rangle}$. Since $[[P^t]^h]_{\langle 1,1\rangle}$ is a permutation matrix, and since $u_{\langle 1\rangle}$ is nonzero, the product of the λ_i must be an eigenvalue for a permutation matrix. □

THEOREM 4.11. *Let U be an inflator associated with an m-partition Π of n. Suppose that $G(U)$ is P-commutative. Let Q be an $m \times m$ permutation matrix such that Π is a Q, P-partition, and let q be the permutation corresponding to Q. Suppose that q decomposes into k disjoint cycles. For each j, let $h(j)$ be the length of the jth cycle in the decomposition of q. Then the following are equivalent:*

(i) *There exists an inflator $\hat{U}$ associated with Π such that $G(U) = G(\hat{U})$ and such that $\hat{U}$ is P-commutative.*

(ii) *There exists a unique, normalized inflator $\hat{U}$ associated with Π such that $G(U) = G(\hat{U})$ and such that $\hat{U}$ is P-commutative.*

(iii) *There exists a constant λ such that for each j with $1 \leqq j \leqq k$, $\lambda^{h(j)}$ equals the product of the λ_r where r runs through the indices in the jth cycle of q, and where the λ_r are given by Lemma 4.7.*

Proof. Suppose that U and $\hat{U}$ are inflators such that $G(U) = G(\hat{U})$. Express U and $\hat{U}$ by $U = uv^t$ and $\hat{U} = \hat{u}\hat{v}^t$ where all of the vectors are strictly nonzero. Note that $G(U) = G(\hat{U})$ if and only if $U_{\langle r,r\rangle} = \hat{U}_{\langle r,r\rangle}$ for $1 \leqq r \leqq m$. Since the vectors are strictly nonzero, this is equivalent to requiring that there exist nonzero constants $\alpha_1, \alpha_2, \cdots, \alpha_m$ such that $\hat{u}_{\langle r\rangle} = \alpha_r u_{\langle r\rangle}$ and $\hat{v}_{\langle r\rangle} = (\alpha_r)^{-1} v_{\langle r\rangle}$ for all r with $1 \leqq r \leqq m$.

By Lemma 2.7, $\hat{U}$ is P-commutative if and only if $P\hat{u} = \gamma\hat{u}$ and $P\hat{v} = \gamma^{-1}\hat{v}$, where γ is a nonzero scalar, that is, if and only if $P^t\hat{u} = \lambda\hat{u}$ and $\hat{v}^t P = \lambda^{-1}\hat{v}^t$ for some nonzero scalar λ, and equivalently, if and only if

$$[P^t]_{\langle q(i),i\rangle}\hat{u}_{\langle i\rangle} = \lambda\hat{u}_{\langle q(i)\rangle} \tag{4.12}$$

and

$$[v_{\langle q(i)\rangle}]^t P = \lambda^{-1}[\hat{v}_{\langle q(i)\rangle}]^t \tag{4.13}$$

hold for all i with $1 \leqq i \leqq m$. Express the subvectors of $\hat{u}$ and $\hat{v}$ in terms of the α_r's and the subvectors of u and v, and then use the fact that $G(U)$ is P-commutative to apply (4.8) and (4.9). Then (4.12) and (4.13) are equivalent to

$$\lambda_i \alpha_i u_{\langle q(i)\rangle} = \lambda \alpha_{q(i)} u_{\langle q(i)\rangle}$$

and

$$\lambda_i^{-1}\alpha_i^{-1} v_{\langle q(i)\rangle} = \lambda^{-1}(\alpha_{q(i)})^{-1} v_{\langle q(i)\rangle}$$

for all i. Since u and v are strictly nonzero, $\hat{U}$ is P-commutative if and only if

$$\lambda_i \alpha_i (\alpha_{q(i)})^{-1} = \lambda \tag{4.14}$$

for all i with $1 \leqq i \leqq m$.

Pick j with $1 \leqq j \leqq k$. Since i and $q(i)$ must lie in the same cycle of q, and since q is a bijection on $\{1, 2, \cdots, m\}$, it follows that if the product of the terms $\lambda_i \alpha_i (\alpha_{q(i)})^{-1}$ is taken over all i in the jth cycle, then the product equals the product of the λ_i as i runs

through the indices of the jth cycle of q. Thus, if $\hat{U}$ is P-commutative, then (4.14) implies that the product of the λ_i as i runs through the indices in the jth cycle of q must be exactly $\lambda^{h(j)}$. Hence (i) implies (iii).

Suppose that condition (iii) is satisfied. In proving (ii), it suffices to show that the α_i can be chosen so that (4.14) holds and so that $\hat{U}$ is normalized.

Consider the jth cycle of q. Without loss of generality, $j = 1$, and the cycle is $(1, 2, 3, \cdots, h)$. Then $q(i) = i + 1 \pmod{h}$ for each i in the cycle. By (iii), it follows that

$$\lambda^h = \lambda_1\lambda_2\lambda_3\cdots\lambda_h.$$

Let α_1 be defined by

$$\alpha_1 = [[v_{\langle 1\rangle}]^* v_{\langle 1\rangle}]^{1/4}[[u_{\langle 1\rangle}]^* u_{\langle 1\rangle}]^{-1/4}.$$

Then

$$\begin{aligned}[\hat{u}_{\langle 1\rangle}]^*\hat{u}_{\langle 1\rangle} &= (\alpha_1)^2[u_{\langle 1\rangle}]^* u_{\langle 1\rangle} = (\alpha_1)^{-2}[v_{\langle 1\rangle}]^* v_{\langle 1\rangle} \\ &= [\hat{v}_{\langle 1\rangle}]^*\hat{v}_{\langle 1\rangle}.\end{aligned}$$

For $2 \leqq i \leqq h$, define α_i by (4.14):

$$\alpha_{i+1} = \lambda^{-1}\lambda_i\alpha_i.$$

It remains to show two things: first, that $\alpha_1 = \lambda^{-1}\lambda_h\alpha_h$, so that (4.14) holds for each i in $(1, 2, 3, \cdots, h)$, and thus that $\hat{U}$ is P-commutative; and second, that $[\hat{u}_{\langle i\rangle}]^*\hat{u}_{\langle i\rangle} = [\hat{v}_{\langle i\rangle}]^*\hat{v}_{\langle i\rangle}$ for each i in $(1, 2, 3, \cdots, h)$ so that the inflator $\hat{U}$ is normalized. To see the first, observe that repeated applications of (4.14) yield

$$\lambda^{-1}\lambda_h\alpha_h = \lambda^{-1}\lambda_h(\lambda^{h-1}\lambda_{h-1}\lambda_{h-2}\cdots\lambda_3\lambda_2\lambda_1\alpha_1) = \alpha_1.$$

To see that the latter is true for $i = 2, 3, \cdots, h$, compute

$$\begin{aligned}[\hat{u}_{\langle i\rangle}]^*\hat{u}_{\langle i\rangle} &= [\alpha_i u_{\langle i\rangle}]^*[\alpha_i u_{\langle i\rangle}] \\ &= [\lambda^{-1}\lambda_{i-1}\alpha_{i-1}u_{\langle i\rangle}]^*[\lambda^{-1}\lambda_{i-1}\alpha_{i-1}u_{\langle i\rangle}] \\ &= |\lambda|^{-2}[\alpha_{i-1}(\lambda_{i-1}u_{\langle i\rangle})]^*[\alpha_{i-1}(\lambda_{i-1}u_{\langle i\rangle})] \\ &= |\lambda|^{-2}[\alpha_{i-1}u_{\langle i-1\rangle}]^*[\alpha_{i-1}u_{\langle i-1\rangle}] \\ &= |\lambda|^{-2}[\hat{u}_{\langle i-1\rangle}]^*[\hat{u}_{\langle i-1\rangle}]\end{aligned}$$

where the second-to-last equality is a consequence of (4.12). Since λ has modulus one from Lemma 4.7, it follows that

$$[\hat{u}_{\langle i\rangle}]^*\hat{u}_{\langle i\rangle} = [\hat{u}_{\langle 1\rangle}]^*\hat{u}_{\langle 1\rangle}$$

for $i = 2, 3, \cdots, h$. A similar argument yields

$$[\hat{v}_{\langle i\rangle}]^*\hat{v}_{\langle i\rangle} = [\hat{v}_{\langle 1\rangle}]^*\hat{v}_{\langle 1\rangle}$$

for $i = 2, 3, \cdots, h$. Finally, $[\hat{u}_{\langle 1\rangle}]^*\hat{u}_{\langle 1\rangle} = [\hat{v}_{\langle 1\rangle}]^* v_{\langle 1\rangle}$ implies $\hat{U}$ is normalized. By Lemma 4.16 of [4], given an inflator U, there is a unique normalized inflator $\hat{U}$ such that $G(U) = G(\hat{U})$. Finally, it is clear that (ii) implies (i). □

Remarks. Since Π is a Q, P-partition, the inflator $\hat{U}$ is P-commutative if and only if it is a Q, P-inflator. The preceding proof shows that if $\hat{U}$ is a normalized, P-commutative inflator, then $\hat{u}$ and $\hat{v}$ can be chosen so that

$$\|\hat{u}_{\langle q(i)\rangle}\| = \|\hat{u}_{\langle i\rangle}\| = \|\hat{v}_{\langle q(i)\rangle}\| = \|\hat{v}_{\langle i\rangle}\|$$

for every i with $1 \leqq i \leqq m$. The proof also contains the following algorithm used to construct $\hat{U}$ from U. Choose a transversal from the disjoint cycles of q. For j in such a transversal, let α_j be given by

$$\alpha_j = [[v_{\langle j \rangle}]^* v_{\langle j \rangle}]^{1/4} [[u_{\langle j \rangle}]^* u_{\langle j \rangle}]^{-1/4}.$$

Now define the remaining α_i via (4.14).

COROLLARY 4.15. *Let U be an inflator. Let P be a permutation matrix. Suppose that $G(U)$ is P-commutative. Let Q be the permutation matrix induced by P, and let q be the permutation corresponding to Q. Then U is a Q, P-inflator if either of the following conditions holds*:

(i) *The cycle decomposition of q consists of a single cycle.*

(ii) *$U = cW$, where $|c| = 1$ and where W is strictly positive.*

Proof. If (i) holds, then condition (iii) of the previous theorem necessarily holds. If (ii) holds, then U can be expressed as $U = [\gamma x][\gamma y]^t$, where x and y are strictly positive, and where $\gamma^2 = c$. If Lemma 4.7 is applied to γx, $[P^t]_{\langle q(i), i \rangle} x_{\langle i \rangle} = \lambda_i x_{\langle q(i) \rangle}$ for each i. Since P is a permutation matrix, and since x is strictly positive, it follows that λ_i must be positive. Then the product of the λ_i as i runs through a cycle in q must be a positive number of modulus one. That is, the product must equal one for each cycle in q. Thus (iii) of Theorem 4.11 holds with $\lambda = 1$. □

5. *ZME*-matrices. A *ZME-matrix* is a matrix all of whose positive (integer) powers are Z-matrices, and all of whose odd, positive powers are irreducible. A *ZMO-matrix* is a *ZME*-matrix all of whose even, positive powers are completely reducible with index of reducibility greater than one. A *ZMA-matrix* is a *ZME*-matrix all of whose positive powers are irreducible. An *MMA-matrix* is a matrix all of whose positive powers are irreducible M-matrices. Note that an *MMA*-matrix is necessarily a *ZME*-matrix.

In [1], Friedland, Hershkowitz, and Schneider prove the following results (see [1, Lemma 3.1; Thms. 3.6, 6.12, 6.18; Cor. 6.25, 6.28]).

THEOREM 5.1. *Let A be a ZME-matrix. Then A has real spectrum and is diagonalizable. Further, if the distinct eigenvalues of A are $\alpha_1, \alpha_2, \cdots, \alpha_k$ labeled so that they satisfy $\alpha_1 < \alpha_2 < \cdots < \alpha_k$, then $|\alpha_1| \leqq \alpha_2$ and α_1 is a simple eigenvalue for A. Finally,*

(i) *A is a ZMO-matrix if and only if $\alpha_1 = -\alpha_2$;*

(ii) *A is a ZMA-matrix if and only if $\alpha_1 > -\alpha_2$;*

(iii) *A is an MMA-matrix if and only if $\alpha_1 \geqq 0$.*

THEOREM 5.2. *Let $\{U_i\}_{i=1}^k$ be a strictly positive inflation sequence. Suppose that U_k is $n \times n$. Let $E_i = G(U_i) \times\times U_{i+1} \times\times \cdots \times\times U_k$ for $1 \leqq i < k$, and let $E_k = G(U_k)$. Let $\Omega = \{E_i\colon 1 \leqq i \leqq k\}$. Then the elements of Ω are pairwise orthogonal, idempotent $n \times n$ real matrices such that $\sum_{i=1}^k E_i = I_n$. Further, A in $\mathcal{M}_n(\mathbb{R})$ is a ZME-matrix with spectrum $\alpha_1, \alpha_2, \cdots, \alpha_k$ satisfying $|\alpha_1| \leqq \alpha_2$ and $\alpha_1 < \alpha_2 < \cdots < \alpha_k$, if and only if A can be expressed as*

$$A = \sum_{i=1}^{k} \alpha_i E_i \tag{5.3}$$

where the E_i form a set Ω arising from an inflation sequence. Finally, if A is a ZME-matrix then A has a unique strictly positive inflation sequence consisting of normalized inflators.

THEOREM 5.4. *Let A be an $n \times n$ ZME-matrix with spectrum $\alpha_1, \alpha_2, \cdots, \alpha_k$ satisfying $|\alpha_1| \leqq \alpha_2$ and $\alpha_1 < \alpha_2 < \cdots < \alpha_k$. Suppose that μ is the multiplicity of α_k in the spectrum of A. Let $m = n - \mu$. Then there exists an m-partition Π of n and a strictly*

positive, normalized inflator U associated with Π *such that*

$$A = C \times\times U + \alpha_k G(U).$$

Further, $G(U)$ *is the spectral projector for* A *corresponding to* α_k, *and* C *is an* $m \times m$ *ZME-matrix with spectrum* $\alpha_1, \alpha_2, \cdots, \alpha_{k-1}$. *Finally, the inflator* U *is uniquely determined by the condition that it is the normalized inflator such that* $G(U)$ *is the spectral projector for* α_k.

6. Main results: *P*-commutative *ZME*-matrices.

THEOREM 6.1. *Let* A *be in* $\mathcal{M}_n(\mathbb{R})$. *Let* P *be an* $n \times n$ *permutation matrix. Suppose that* A *is a ZME-matrix. The following are equivalent*:

(i) A *is a P-commutative matrix.*

(ii) *Each of the spectral projectors of* A *is a real, P-commutative matrix.*

(iii) A *has a normalized, strictly positive inflation sequence* $\{U_i\}_{i=1}^{k}$ *for which there exists a sequence of permutation matrices* $\{P_i\}_{i=1}^{k}$ *such that for* $i \geqq 2$, U_i *is a* P_{i-1}, P_i*-commutative inflator, and such that* $P_k = P$.

Proof of (i) $\Rightarrow$ (iii). Use induction on the number of distinct eigenvalues in the spectrum of A. Let $\alpha_1, \alpha_2, \cdots, \alpha_k$ be the distinct eigenvalues of A, in order of increasing magnitude. If $k = 1$, then by Theorem 5.2, A has an inflation sequence consisting of $U_1 = [0]$ which is clearly P-commutative.

Suppose that $k > 1$. Since A is a centrosymmetric ZME-matrix, it follows that the spectral projectors for A are P-commutative matrices by Theorem 2.4. By Theorem 5.4, A can be expressed as $C \times\times U_k + \alpha_k G(U_k)$, where U_k is the normalized inflator which is uniquely determined by the condition that $G(U_k)$ is the spectral projector for α_k. Since $G(U_k)$ is the spectral projector for α_k, it follows that $G(U_k)$ is a P-commutative matrix. By Corollary 4.15, it follows that $G(U_k) = G(\hat{U}_k)$, where $\hat{U}_k$ is a normalized, P_{k-1}, P-commutative inflator. By uniqueness, $U_k = \hat{U}_k$. Since U_k is strictly positive, it follows by Lemma 3.2 that C is a P_{k-1}-commutative ZME-matrix. By induction, C has an inflation sequence $\{U_i\}_{i=1}^{k-1}$ consisting of normalized, P_{i-1}, P_i-commutative inflators. Since the multiplicity of α_k in the spectrum of A is $(n - m)$, it follows that $m < n$, and thus $\{U_i\}_{i=1}^{k}$ is an inflation sequence for A with the desired properties.

Proof of (iii) $\Rightarrow$ (ii). Suppose that A is given as in Theorem 5.2. The matrices E_i defined in Theorem 5.2 are precisely the spectral projectors of A. Thus it suffices to show that each E_i is a real, P-commutative matrix. Since A has real spectrum, it follows from Corollary 2.6 that the spectral projectors are real matrices. By Lemma 4.4, it follows that each E_i is P-commutative.

Proof of (ii) $\Rightarrow$ (i). Since A is a linear combination of its spectral projectors, and since the P-commutative matrices form an algebra, the result is clear. □

Remark. Since powers of P-commutative matrices are P-commutative, it follows that if A is a P-commutative ZME-matrix, then every positive integer power of A is a P-commutative Z-matrix.

COROLLARY 6.2. *Let* A *be in* $\mathcal{M}_n(\mathbb{R})$. *Suppose that* A *is a ZME-matrix. The following are equivalent*:

(i) A *is a centrosymmetric matrix.*

(ii) *Each of the spectral projectors of* A *is a real, centrosymmetric matrix.*

(iii) A *has a normalized, strictly positive inflation sequence* $\{U_i\}_{i=1}^{k}$ *for which there exists a sequence of permutation matrices* $\{P_i\}_{i=1}^{k}$ *such that for* $i \geqq 2$, U_i *is a* P_{i-1}, P_i*-commutative inflator, and such that* $P_k = J_n$.

Note that for centrosymmetric matrices, (iii) does not hold with every P_i equal to the permutation matrix J of the appropriate order. Indeed, as the following example

shows, there exist ZME-matrices A such that A is centrosymmetric, but such that every inflation sequence for A contains an inflator which is not centrosymmetric.

Example. Let $U_1 = [0]$. Let $U_2 = uu^t$, where $u = 1/5(3\ 4)^t$. Let $U_3 = vv^t$, where $v = (1/\sqrt{2}\ 3/5\ 4/5\ 1/\sqrt{2})^t$. Let $\Pi_2 = \{1, 2\}$. Let $\Pi_3 = \{1, 4\}, \{2, 3\}$. Observe that if Π_3 is a Q, J_4-commutative partition, then Q must be the 2×2 permutation matrix I_2 since the permutation p corresponding to J_4 has cycle structure $(1, 4)(2, 3)$. Let A be any ZME-matrix for which $\{U_i\}_{i=1}^3$ is an inflation sequence. It is easily verified that A is centrosymmetric. Suppose that $\{W_i\}_{i=1}^3$ is another inflation sequence for A. It can be shown that W_2 must be positive diagonally similar to V_2. Since such a transformation preserves the diagonal entries of U_2, V_2 cannot be centrosymmetric. Alternatively, note that $G(W_3) = G(U_3)$. From the structure of $G(U_3)$, it follows that the 2-partition of 4 that corresponds to V_3 would have to be either $\{1, 4\}, \{2, 3\}$ or $\{2, 3\}, \{1, 4\}$. Neither of these is a J_2, J_4-partition.

Acknowledgments. The author thanks Professor James Weaver of the University of Western Florida for introducing him to the subject of centrosymmetric matrices, and for raising the questions that led to the original version of this paper, which treated centrosymmetric matrices. The author also thanks an anonymous referee for suggesting that the paper be extended to the subject of P-commutativity, and for suggesting the introduction of Lemmas 2.7 and 4.5 (in their centrosymmetric versions) to shorten the proofs of Theorems 4.1 and 4.11.

REFERENCES

[1] S. FRIEDLAND, D. HERSHKOWITZ, AND H. SCHNEIDER, *Matrices whose powers are M-matrices or Z-matrices*, Trans. Amer. Math. Soc., 300 (1987), pp. 343–366.

[2] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1959.

[3] I. J. GOODE, *The inverse of a centrosymmetric matrix*, Technometrics, 12 (1970), pp. 925–928.

[4] R. G. KHAZANIE, *An indication of the asymptotic nature of the Mendelian Markhov process*, J. Appl. Probab., 5 (1968), pp. 350–356.

[5] P. LANCASTER, *Theory of Matrices*, Academic Press, New York, 1969.

[6] J. R. WEAVER, *Centrosymmetric (cross-symmetric) matrices, their basic properties, eigenvalues and eigenvectors*, MAA Monthly, 92 (1985), pp. 711–717.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 3, July 1988

014

ALGORITHMS FOR MATRIX TRANSPOSITION ON BOOLEAN N-CUBE CONFIGURED ENSEMBLE ARCHITECTURES*

S. LENNART JOHNSSON† AND CHING-TIEN HO‡

Abstract. In a multiprocessor with distributed storage the data structures have a significant impact on the communication complexity. In this paper we present a few algorithms for performing matrix transposition on a Boolean n-cube. One algorithm performs the transpose in a time proportional to the lower bound both with respect to communication start-ups and to element transfer times. We present algorithms for transposing a matrix embedded in the cube by a binary encoding, a *binary-reflected* Gray code encoding of rows and columns, or combinations of these two encodings. The transposition of a matrix when several matrix elements are identified to a node by *consecutive* or *cyclic* partitioning is also considered and lower bound algorithms given. Experimental data are provided for the Intel iPSC and the Connection Machine.

Key words. matrix transpose, Boolean cubes, personalized communication, routing, data encoding

AMS(MOS) subject classifications. 65F30, 68P99, 68Q20, 68Q25

1. Introduction.

Matrix transposition is a permutation frequently performed in linear algebra. It is useful in the solution of systems of linear equations by a variety of techniques. For instance, the solution of partial differential equations by the Alternating Direction Method (ADM) is typically carried out by transposing the data between the solution phases in the different directions. Such data transposition may also be beneficial with respect to performance for the ADM on Boolean n-cube configured architectures, even though multidimensional arrays can be embedded in Boolean cubes preserving proximity [12], [13]. Another example where data transposition may be advantageous is the solution of Poisson's problem by the Fourier Analysis Cyclic Reduction (FACR) method. Matrix transposition can also be used to realize arbitrary permutations [21], [20].

In this paper we focus on matrix transposition on Boolean n-cube architectures. The transpose can be formed recursively as described in [19], [1], [8], [15]. H. S. Stone [19] describes a mapping to shuffle-exchange networks for the case with one matrix element per node. We consider the case with multiple matrix elements per node and focus on the pipelining of communication operations and the optimal use of the communication bandwidth of the Boolean n-cube. In [8], [9] we describe and analyze the complexity of a transpose algorithm for a two-dimensional mesh and present a few algorithms for the transposition of matrices embedded in the cube by binary or Gray code encoding of the row and column indices. In this paper we present a transpose algorithm that is of lower complexity in the case of concurrent communication on multiple ports, and present experimental data for the Intel iPSC and the Connection Machine [3].

We first introduce the notation and data structures used in this study, then present algorithms for the transpose operation for one-dimensional and two-dimensional partitionings. Implementation issues particular to the actual machines used, but important

* Received by the editors October 21, 1986; accepted for publication November 30, 1987. This research was partly supported by the Office of Naval Research under contracts N00014-84-K-0043 and N00014-86-K-0564.

† Department of Computer Science and Electrical Engineering, Yale University, New Haven, Connecticut 06520; Thinking Machine Corp., 245 First Street, Cambridge, Massachusetts 02142.

‡ Department of Computer Science, Yale University, New Haven, Connecticut 06520.

for the interpretation of the experimental results presented, are addressed after the description of the algorithms. A summary and conclusion follows.

2. Preliminaries. Let A be a $P \times Q$ matrix. Throughout the paper, we assume that $P = 2^p$ and $Q = 2^q$. The number of bits required for the encoding of the matrix elements is $m = p + q$. The transpose A^T of A is defined by the relation $a^T(u, v) = a(v, u)$, where $a^T(u, v)$ is the element in row u and column v of A^T, and $a(u, v)$ is the element of A in row u and column v. Let the binary encoding of u be $(u_{p-1}u_{p-2}\cdots u_0)$ and the binary encoding of v be $(v_{q-1}v_{q-2}\cdots v_0)$. Then the address of element $a(u, v)$ is naturally defined to be $(u_{p-1}u_{p-2}\cdots u_0 v_{q-1}v_{q-2}\cdots v_0) = (w_{m-1}w_{m-2}\cdots w_0)$, or $(u||v) = w$ for short, where "||" is the concatenation operator for binary numbers.

DEFINITION 1. The matrix transposition operation is the permutation $\mathrm{loc}(u_{p-1}u_{p-2}\cdots u_0 v_{q-1}v_{q-2}\cdots v_0) \leftarrow \mathrm{loc}(v_{q-1}v_{q-2}\cdots v_0 u_{p-1}u_{p-2}\cdots u_0)$, where $\mathrm{loc}(w)$ is the memory location of element w.

Note that we arbitrarily assumed that the p highest-order dimensions are used for the encoding of row indices. We use this assumption throughout this paper, but any other subset of p dimensions could have been used.

With the assumption above, the p highest-order dimensions encode row indices before the transposition and the q highest-order dimensions encode column indices after the transposition. A vector transposition requires no data movement. For the matrix transposition it is sometimes appropriate to consider a square array of $2\max(p, q)$ dimensions.

DEFINITION 2. A $P \times Q$ matrix with $P > Q$ is extended to a square matrix by introducing *virtual elements* corresponding to $P - Q$ columns. The extension is made similarly if $P < Q$.

The extension can be made by adding columns corresponding to high- or low-order dimensions of the column address space, or by mixing columns of virtual elements with columns of real elements. Whichever alternative is preferable depends on the particular transposition algorithm, and data assignment scheme (described later).

DEFINITION 3. A *shuffle* operation, sh^1 on a set of elements $\mathcal{W}$ with addresses w, $w \in \{0, 1, \cdots, 2^m - 1\}$ encoded in binary representation $(w_{m-1}w_{m-2}\cdots w_0)$ is a permutation defined by a one step *left cyclic shift*, $\mathrm{loc}(w_{m-1}w_{m-2}\cdots w_0) \leftarrow \mathrm{loc}(w_{m-2}w_{m-3}\cdots w_0 w_{m-1})$, $w \in \{0, 1, \cdots, 2^m - 1\}$. An *unshuffle* operation, sh^{-1} is defined by a one-step right cyclic shift. $sh^k = sh \circ sh^{k-1}$ is a k step left cyclic shift.

Clearly, $sh^1 \circ sh^{-1} = I$, where I is the identity operator. Also, $sh^k(w) = sh^{-(m-k)}(w)$.

LEMMA 2.1. *Let A be a $2^p \times 2^q$ matrix. $A^T \leftarrow sh^p A$, or $A^T \leftarrow sh^{-q} A$.*

COROLLARY 2.2. *On a shuffle-exchange network of $N = 2^n$ nodes, $n = p + q$, and bidirectional communication links, the matrix transposition requires at most $\min(p, q)$ communication steps.*

A shuffle-exchange network has all the connections corresponding to the sh^1 operation, and connections from every even node to the succeeding odd node.

DEFINITION 4. Let $w = (w_{m-1}w_{m-2}\cdots w_0)$ and $z = (z_{m-1}z_{m-2}\cdots z_0)$. Then $\mathrm{Hamming}(w, z) = \sum_{i=0}^{m-1}(w_i \oplus z_i)$, where $\oplus$ is the exclusive or operation.

LEMMA 2.3. *For m even there exists at least one w such that* $\mathrm{Hamming}(w, sh^1 w) = m$, *and for m odd* $\mathrm{Hamming}(w, sh^1 w) = m - 1$. *In general, for k shuffles*

$$\max_w \mathrm{Hamming}(w, sh^k w) = \begin{cases} m, & \frac{m}{\gcd(m,k)} \text{ is even,} \\ m - \gcd(m, k), & \frac{m}{\gcd(m,k)} \text{ is odd.} \end{cases}$$

Proof. For m even, let $w = (0101\cdots01)$. Then Hamming$(w, sh^1w) = m$. For m odd, let $w = (0101\cdots010)$. Then Hamming$(w, sh^1w) = m-1$. Note that w and sh^1w contain the same numbers of 0's and 1's. Since one of them is odd, the Hamming distance between w and sh^1w is at most $m-1$. For k shuffles, the bits can be divided into $\gcd(m,k)$ groups of bit strings of length $\frac{m}{\gcd(m,k)}$. The lemma follows. □

COROLLARY 2.4. *For m even* $\max_w$ Hamming$(w, sh^{m/2}w) = m$.

LEMMA 2.5. *For* $0 \le k < m$, $\max_w$ Hamming$(w, sh^kw) \ge k$.

Proof. Since $m > k$ we have

$$\frac{m}{\gcd(m,k)} > \frac{k}{\gcd(m,k)} \quad \text{or} \quad \frac{m}{\gcd(m,k)} \ge 1 + \frac{k}{\gcd(m,k)}.$$

This means $m - \gcd(m,k) \ge k$. Lemma 2.3 completes the proof. □

DEFINITION 5. Let $x = (x_{n-1}x_{n-2}\cdots x_0)$, $x_i \in \{0,1\}$, for all $i \in \{0,1,\cdots,n-1\}$ be the address of a node in a Boolean n-cube. Then node x is connected to nodes in the set $\{(x_{n-1}x_{n-2}\cdots\overline{x_i}\cdots x_0) | \text{ for all } i \in \{0,1,\cdots,n-1\}\}$.

A Boolean n-cube has $N = 2^n$ nodes, and each node n neighbors. The diameter is n and the number of links is $\frac{1}{2}nN$. There exist n paths between any pair of nodes (x,y). Of these paths Hamming(x,y) paths are of length Hamming(x,y) and n − Hamming(x,y) paths are of length Hamming$(x,y) + 2$ [18]. We will use this property in devising transposition algorithms with multiple paths between source and destination processors for minimization of the data transfer time.

LEMMA 2.6. *Matrix transposition on a Boolean n-cube requires at least as many communication steps as the transposition on a shuffle-exchange network.*

Lemma 2.6 is immediate from Lemma 2.5.

In general, the number of matrix elements may be larger than the number of processors, and several matrix elements must be allocated to the storage of individual processors. We assume that the matrix elements are distributed evenly among the processors. For $n \le \max(p,q)$ there is a choice between one- and two-dimensional partitioning. For either kind of partitioning the matrix elements can be assigned to processors *cyclicly*, or *consecutively* [8], [9], or by a *combined* assignment scheme.

DEFINITION 6. In a one-dimensional *cyclic* partitioning on N processors, row u (column v) is assigned to processor $u \bmod N$ ($v \bmod N$) and in a one-dimensional *consecutive* partitioning row u (column v) is assigned to processor $\lfloor u/\lceil \frac{P}{N} \rceil \rfloor$ ($\lfloor v/\lceil \frac{Q}{N} \rceil \rfloor$).

COROLLARY 2.7. *In an n-cube the n lowest-order bits of the binary encoded row (column) index determines the processor to which a row (column) is assigned in the cyclic partitioning. In the consecutive assignment the n highest-order bits determines the processor assignment, if the number of rows (columns) is a power of* 2.

The dimensions that are of higher- (lower-) order than the real processor address field are used for cyclic (consecutive) assignment. The notions of cyclic and consecutive assignment are relative to a given real processor address field.

In the two-dimensional partitioning we let $N_r = 2^{n_r} \le P$ denote the number of partitions in the row direction and let $N_c = 2^{n_c} \le Q$ denote the number of partitions in the column direction. The total number of partitions is $N_r \times N_c \le N$ ($n_r + n_c \le n$). In the cyclic partitioning matrix element (u,v) is assigned to partition ($u \bmod N_r, v \bmod N_c$) and in the consecutive partitioning it is assigned to partition $(\lfloor u/\lceil \frac{P}{N_r} \rceil \rfloor, \lfloor v/\lceil \frac{Q}{N_c} \rceil \rfloor)$ (Fig. 2). For a matrix partitioned by cyclic assignment the n_r lowest-order bits of the matrix row index determines the processor row index. Analogously, the n_c lowest-order bits of the matrix column index determines the processor column index. In consecutive storage, the n_r highest-order bits in the

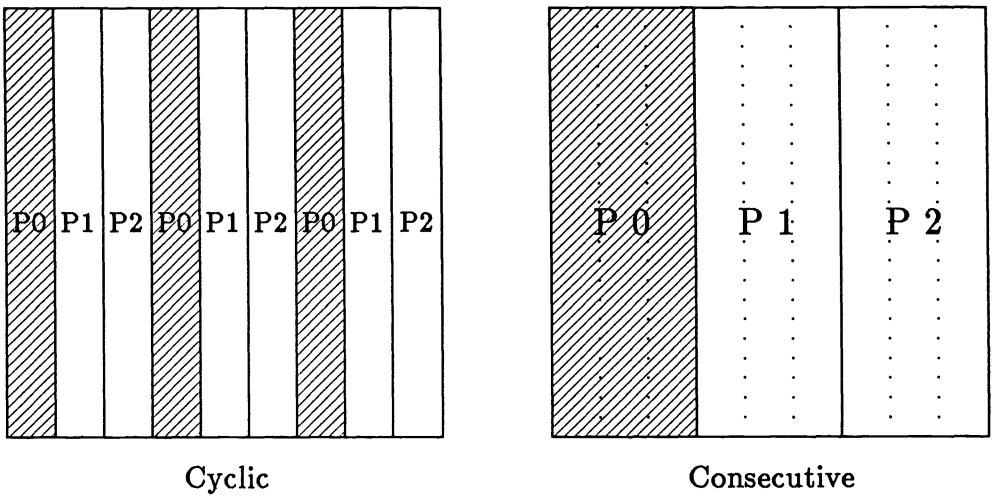


FIG. 1. *Cyclic and consecutive one-dimensional partitioning.*

matrix row index determine the processor row index and the n_c highest-order bits of the column index determine the processor column index, since P and Q are powers of two.

The cyclic and consecutive assignment schemes are illustrated in Figs. 1 and 2 with respect to the matrix elements.

DEFINITION 7. The part of the address field that does not correspond to real processors defines *virtual processors.*

The cyclic and consecutive assignment schemes with respect to the address space is as follows: One-dimensional cyclic column partitioning:

$$(\underbrace{u_{p-1}u_{p-2}\cdots u_0\ v_{q-1}v_{q-2}\cdots v_{n_c}}_{vp}\ \underbrace{v_{n_c-1}\cdots v_0}_{rp}).$$

One-dimensional consecutive column partitioning:

$$(\underbrace{u_{p-1}u_{p-2}\cdots u_0}_{vp}\ \underbrace{v_{q-1}v_{q-2}\cdots v_{q-n_c}}_{rp}\ \underbrace{v_{q-n_c-1}\cdots v_0}_{vp}).$$

For the cyclic two-dimensional assignment the address field is partitioned as

$$(\underbrace{u_{p-1}u_{p-2}\cdots u_{n_r}}_{vp}\ \underbrace{u_{n_r-1}\cdots u_0}_{rp}\ \underbrace{v_{q-1}v_{q-2}\cdots v_{n_c}}_{vp}\ \underbrace{v_{n_c-1}\cdots v_0}_{rp}),$$

and for the consecutive assignment the address field is partitioned as

$$(\underbrace{u_{p-1}u_{p-2}\cdots u_{p-n_r}}_{rp}\ \underbrace{u_{p-n_r-1}\cdots u_0}_{vp}\ \underbrace{v_{q-1}v_{q-2}\cdots v_{q-n_c}}_{rp}\ \underbrace{v_{q-n_c-1}\cdots v_0}_{vp}),$$

where vp denotes the dimensions of the address space used for virtual processor addresses and rp denotes the dimensions used for real processor addresses. The number of dimensions used for the consecutive, or cyclic mapping is $m - n_c$ (or $m - n_r$) in

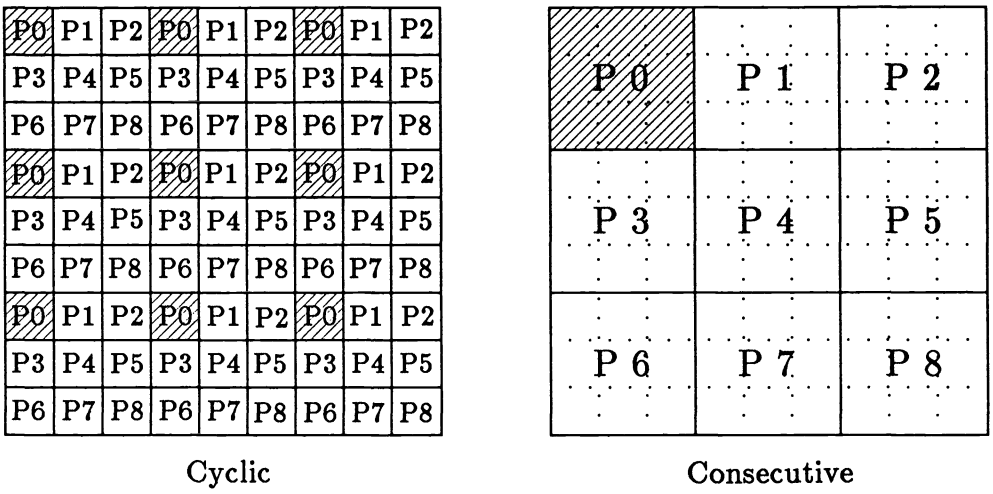


FIG. 2. *Cyclic and consecutive two-dimensional partitioning.*

the one-dimensional case and $m - n_r - n_c$ in the two-dimensional case. For column partitioning, $n_r = 0$, $0 \leq n_c \leq n$. For row partitioning, $n_c = 0$, $0 \leq n_r \leq n$.

The *cyclic* and *consecutive* storage forms are two extreme cases of real processor assignment. We refer to the general case as *combined* assignment. Any subset of dimensions of the address space can be used for real processor addresses. As an example of combined assignment we consider the storage of a banded matrix for the equation solvers in [7], [11]. The nonzero elements of the matrix, the right-hand sides, and the solution vectors can be stored in a rectangular array by conventional row/column storage of the matrix, or by row/diagonal organization. Here we do not discuss the techniques for band matrix storage and their consequences for the solution procedure. For illustration we simply assume that the relevant elements are stored in an array of $P = 2^p$ rows and $Q = 2^q$ columns. Then, for a two-dimensional partitioning with 2^{n_c} processors in both the row and column directions, blocks of size $2^{q-n_c} \times 2^{q-n_c}$ elements may be stored in the same processor, and blocks assigned cyclically with respect to the row addresses, i.e., the address field is partitioned as

$$(\underbrace{u_{p-1}u_{p-2}\cdots u_q}_{vp}\ \underbrace{u_{q-1}\cdots u_{q-n_c}}_{rp}\ \underbrace{u_{q-n_c-1}\cdots u_0}_{vp}\ \underbrace{v_{q-1}\cdots v_{q-n_c}}_{rp}\ \underbrace{v_{q-n_c-1}\cdots v_0}_{vp}).$$

The total number of real processor dimensions is $2n_c$. For the row assignment the n_c contiguous dimensions of the address field used for real processor addresses divides the address space into two parts: $q - n_c$ dimensions used for consecutive assignment, and $p - q$ dimensions used for cyclic assignment. For the concurrent elimination of multiple vertices the matrix is partitioned into S block rows. With $S = 2^s$ the s highest-order bits of the matrix row addresses are used for real processor addresses. With the previous assignment for each such block the address field is partitioned as

$$(\underbrace{u_{p-1}\cdots u_{p-s}}_{rp}\ \underbrace{u_{p-s-1}\cdots u_q}_{vp}\ \underbrace{u_{q-1}\cdots u_{q-n_c}}_{rp}\ \underbrace{u_{q-n_c-1}\cdots u_0}_{vp}\ \underbrace{v_{q-1}\cdots v_{q-n_c}}_{rp}\ \underbrace{v_{q-n_c-1}\cdots v_0}_{vp}).$$

Hence, in this case the dimensions used for real processor addresses forms two fields. The number of dimensions for real processors in the row direction is $s + n_c$, and

the total number of real processor dimensions is $s + 2n_c$. The notions of cyclic and consecutive partitioning are now conditioned on the part of the real processor address fields.

We now turn to the communication required for matrix transposition. Consider a one-dimensional partitioning such that $p = q > n_c = n$ and cyclic partitioning by columns before the transposition. Then every processor sends 2^{m-2n} elements to every other processor. *All-to-all personalized communication* [5], [14] is required. To see this fact, note that there are 2^{m-n} virtual processors per real processor, and that the address field prior to the transposition is partitioned as

$$(\underbrace{u_{p-1}u_{p-2}\cdots u_0\ v_{q-1}v_{q-2}\cdots v_n}_{vp}\ \underbrace{v_{n-1}\cdots v_0}_{rp}).$$

After the transposition the partitioning is

$$(\underbrace{v_{q-1}v_{q-2}\cdots v_0\ u_{p-1}u_{p-2}\cdots u_n}_{vp}\ \underbrace{u_{n-1}\cdots u_0}_{rp}),$$

which, in the original address field, is

$$(\underbrace{u_{p-1}u_{p-2}\cdots u_n}_{vp}\ \underbrace{u_{n-1}\cdots u_0}_{rp}\ \underbrace{v_{q-1}v_{q-2}\cdots v_0}_{vp}).$$

Hence, the row address field in the initial allocation is partitioned into 2^n partitions for each column, and each such partition sent to a unique processor for the matrix transposition. The address fields for real processors before the transposition and after the transposition are disjoint.

If $q < n \leq p$ and the initial assignment is by columns, then only 2^q processors are used before the transposition, but all 2^n processors are used after the transposition. The number of virtual processors per real processor before the transposition is 2^p, and after the transposition is 2^{m-n}. The row address field is divided into 2^n partitions. The address fields for real processors before and after the transposition are disjoint. The transposition is accomplished by all 2^q processors holding matrix elements sending a unique set of data to each of the 2^n processors. The communication is *some-to-all personalized communication.* The reverse operation is *all-to-some personalized communication.* In the extreme case such as transposing a vector, it is *one-to-all* or *all-to-one personalized communication.* In a two-dimensional partitioning with the same number of processors assigned to rows and columns, and the same assignment scheme (cyclic or consecutive) for rows and columns, the address fields for real processors before and after the transposition are the same. The communication is between distinct pairs of processors.

One of the reasons for not using all processors before or after a transposition is that the number of dimensions for the row or column address field is smaller than the number of processors dimensions assigned to that address field. *Virtual elements* can be introduced to simplify the analysis. *Virtual processors* define local storage addresses.

Let $\mathcal{R}$ be the set of dimensions used for real processors, and $\mathcal{V}$ the set of dimensions used for virtual processors: $\mathcal{R} = \{d_i | i = 0, 1, \cdots, rp - 1\}$ and $\mathcal{V} = \{d'_i | i = 0, 1, \cdots, vp - 1\}$, where $d_i, d'_i \in \{0, 1, \cdots, m\}$. Furthermore $\mathcal{R} \cap \mathcal{V} = \phi$ and $\mathcal{R} \cup \mathcal{V} = \{0, 1, \cdots m - 1\}$. The number of dimensions used for real processor addresses is

TABLE 1

The processor address for matrix element $(u_{p-1}u_{p-2}\cdots u_0, v_{q-1}v_{q-2}\cdots v_0)$ *with consecutive and cyclic encodings.*

Enc./Part.	Consecutive	Cyclic
binary, row	$(u_{p-1}u_{p-2}\cdots u_{p-n})$	$(u_{n-1}u_{n-2}\cdots u_0)$
binary, column	$(v_{q-1}v_{q-2}\cdots v_{q-n})$	$(v_{n-1}v_{n-2}\cdots v_0)$
Gray, row	$(G(u_{p-1}u_{p-2}\cdots u_{p-n}))$	$(G(u_{n-1}u_{n-2}\cdots u_0))$
Gray, column	$(G(v_{q-1}v_{q-2}\cdots v_{q-n}))$	$(G(v_{n-1}v_{n-2}\cdots v_0))$

TABLE 2

The processor address for matrix element $(u_{p-1}u_{p-2}\cdots u_0, v_{q-1}v_{q-2}\cdots v_0)$ *with two examples of combined encoding.*

Enc./Part.	Combined	
	Contiguous	Noncontiguous
binary, row	$(u_{p-i}u_{p-i-1}\cdots u_{p-i-n+1})$	$(u_{p-1}\cdots u_{p-s}u_{n-s-1}\cdots u_0)$
binary, column	$(v_{q-i}v_{q-i-1}\cdots v_{q-i-n+1})$	$(v_{q-1}\cdots v_{q-s}v_{n-s-1}\cdots v_0)$
Gray, row	$(G(u_{p-i}u_{p-i-1}\cdots u_{p-i-n+1}))$	$(G(u_{p-1}\cdots u_{p-s})G(u_{n-s-1}\cdots u_0))$
Gray, column	$(G(v_{q-i}v_{q-i-1}\cdots v_{q-i-n+1}))$	$(G(v_{q-1}\cdots v_{q-s})G(v_{n-s-1}\cdots v_0))$

$|\mathcal{R}| = rp$, and the number of dimensions used for virtual processors is $|\mathcal{V}| = vp$ $(rp + vp = m)$. Denote the set of dimensions of the matrix encoding assigned to real processors before the transposition by $\mathcal{R}_b$ and the set of matrix dimensions used for real processors after the transposition by $\mathcal{R}_a$. Let $\mathcal{I} = \mathcal{R}_b \cap \mathcal{R}_a$. Clearly, for any one-dimensional partitioning $\mathcal{I} = \phi$.

So far we have assumed that the matrix elements are embedded in the set of processors by a binary encoding. Such an embedding does not preserve proximity. A *binary-reflected Gray code* [16] encoding of row and column indices preserves adjacency. This code is referred to as the Gray code in the following and the encoding of w is $G(w)$. The conversion from one kind of encoding to the other can be accomplished in $n - 1$ routing steps with additional local data rearrangement. The paths in the routing can be made to be edge-disjoint [8].

Adjacency is of no concern for virtual processor addresses in a storage with uniform access time, but may be of significance for interprocessor communication, in particular for Boolean cube configured multiprocessors. It is possible to restrict the Gray code encoding to the real processor address field. For instance, in the consecutive assignment the stripes/blocks can be assigned to processors by a Gray code encoding, while the elements within the stripes/blocks are ordered in the binary order.

If we consider binary and Gray code encoding of the processor address field, and consecutive, cyclic, or combined assignment with a consecutive or split address field, a total of 16 matrix embeddings result for a one-dimensional partitioning. The conversions between any two of the 16 assignment schemes are equivalent, i.e., *all-to-all personalized communication*, in terms of the global communication, if $\mathcal{I} = \phi$ and $|\mathcal{R}_a| = |\mathcal{R}_b| = |\mathcal{R}|$. Table 2 shows the real address fields and their encoding in terms of the matrix dimensions for *consecutive* and *cyclic* assignments. Table 2 shows the encodings for two examples of *combined* assignment. The general case, for which n arbitrarily chosen dimensions are used for real processor addresses, is treated in [4].

For the architecture we assume that the communication is packet oriented with a communications overhead τ, a transmission time per element t_c, and a maximum packet size of B_m elements. A communications overhead is incurred for each communications link traversed. For a bit-serial architecture, such as the Connection Machine, the overhead is only incurred once through pipelining. With the operating system for the Intel iPSC on which our experiments were carried out $\tau \approx 5$ *m*sec, $t_c \approx 1$ μsec/byte and $B_m = 1$ *k*bytes. For the algorithm description and analysis we consider two cases with respect to communication capabilities: communication restricted to one port at a time, *one-port* communication, and concurrent communication on all ports, *n-port* communication. *One-port* communication is a good approximation of the capabilities of the Intel iPSC. Furthermore, we assume bidirectional communication, i.e., that a processor can send and receive data concurrently on the same port. Therefore, one send *or* one receive operation takes the same time as one exchange operation of two adjacent nodes through the same link for both *one-port* and *n-port* communications.

3. Generic algorithms.

3.1. One-to-all personalized communication. In [14] we devised and analyzed algorithms for *one-to-all* and *all-to-all personalized communication. One-to-all personalized communication* can be performed in a time within a factor of two of the lower bound by routing according to a *Spanning Binomial Tree* (SBT) with *one-port* communication [17], [2], [5]. Before the communication the source node holds all PQ data elements. After the communication, every processor holds PQ/N data elements. The communication time for SBT routing and scheduling all data for a subtree at once [5] is $T = (1 - \frac{1}{N})PQt_c + \sum_{i=1}^{n} \lceil \frac{PQ}{2^i B_m} \rceil \tau$, which is minimized for $B_m \geq \frac{PQ}{2}$. $T_{\min} = (1 - \frac{1}{N})PQt_c + n\tau$. The lower bound $T_{l\,b} \geq \max((1 - \frac{1}{N})PQt_c, n\tau) \geq \frac{1}{2}((1 - \frac{1}{N})PQt_c + n\tau)$.

With *n-port* communication routing according to an SBT results in a time complexity of an order higher than the lower bound. Half of the nodes of an SBT are in one of the subtrees of the root node, and the minimum transmission time is $\frac{1}{2}PQt_c$. The lower bound for *n-port* communication is $T_{l\,b} \geq \max(\frac{1}{n}(1 - \frac{1}{N})PQt_c, n\tau) \geq \frac{1}{2}(\frac{1}{n}(1 - \frac{1}{N})PQt_c + n\tau)$. One routing strategy optimal within a small constant factor is to use a *Spanning Balanced n-Tree* (SBnT) [5], [14], [6]. The communication time for SBnT routing and scheduling data for each subtree in a reverse breadth-first order is

$$\begin{aligned} T &= \sum_{i=1}^{n} \left(\frac{1}{n} \binom{n}{i} \frac{PQ}{N} t_c + \left\lceil \frac{1}{n} \binom{n}{i} \frac{PQ}{B_m N} \right\rceil \tau \right) \\ &= \frac{1}{n} \left(1 - \frac{1}{N}\right) PQt_c + \sum_{i=1}^{n} \left(\left\lceil \frac{1}{n} \binom{n}{i} \frac{PQ}{B_m N} \right\rceil \tau \right), \end{aligned}$$

which has a minimum of

$$T_{\min} = \frac{1}{n} \left(1 - \frac{1}{N}\right) PQt_c + n\tau \quad \text{for} \quad B_m \geq \max_{\forall i} \binom{n}{i} \frac{1}{n} \frac{PQ}{N} \approx \sqrt{\frac{2}{\pi}} \frac{PQ}{n^{3/2}}.$$

The speed-up of the transmission time of the SBnT routing over the SBT routing is a factor of $\frac{n}{2}$. The maximum packet size is reduced approximately by a factor of n.

In the SBnT routing the node set is divided into n approximately equal sets. An alternative routing for *n-port* communication, is to divide the data set ($\frac{PQ}{N}$) for each node into n equal parts and route the parts according to SBT's *rotated* with respect to each other if $\frac{PQ}{N} \bmod n = 0$.

DEFINITION 8. A graph is *rotated* with respect to another graph if all its addresses are obtained through the same number of shuffle operations, sh^k for some k, of the addresses of the other graph.

DEFINITION 9. A graph is a *reflection* of another graph if all its addresses are obtained through a bit-reversal of the addresses of the other graph.

Note that in the case of the SBT, a reflected SBT can be obtained by complementing trailing zeros instead of leading zeros. The minimum time for *one-to-all personalized communication* using n distinctly rotated spanning binomial trees and scheduling data for each subtree in a reverse breadth-first order is $T_{\min} = \frac{1}{n}(1 - \frac{1}{N})PQt_c + n\tau$ [5]. This complexity is of the same order as that of the lower bound. The minimum time is achieved for $B_m \geq \sqrt{\frac{2}{\pi}}\frac{PQ}{n^{3/2}}$. A similar algorithm of the same complexity was also derived independently by Stout and Wager [21], [20].

For $\frac{PQ}{N} = k < n$ the SBnT routing has a lower time complexity for element transfers. For k SBT's the transfer time for optimally rotated spanning binomial trees is

$$(2^n - 1)\frac{2^{n/k-1}}{2^{n/k} - 1}\frac{PQ}{N}t_c$$

and for optimally reflected and rotated spanning binomial trees the minimum transfer time with concurrent communication on all ports is

$$(2^n - 1)\frac{2^{2n/k-1} + 1}{2^{2n/k} - 1}\frac{PQ}{N}t_c.$$

For $k = 2$ reflection yields a maximum of $\frac{N}{2} + 1$ element transfers over any edge (and a minimum of $\sqrt{2N}$). Rotation yields a maximum of $\frac{N}{2} + \sqrt{\frac{N}{2}}$ element transfers over any edge. For $k = 2$ the optimum rotation is by $\frac{n}{2}$ steps. In general, the optimum rotation is by $\frac{n}{k}$ steps for $\frac{PQ}{N} = k < n$, if n is a multiple of k.

3.2. All-to-all personalized communication. For *all-to-all personalized communication* a simple exchange algorithm scanning through the dimensions of the cube for *one-port* communication requires a time

$$T = n\frac{PQ}{2N}t_c + n\left\lceil\frac{PQ}{B_m 2N}\right\rceil\tau,$$

which has the minimum

$$T_{\min} = n\left(\frac{PQ}{2N}t_c + \tau\right) \quad \text{for} \quad B_m \geq \frac{PQ}{2N}$$

[17], [8], [15], [14], [2]. In each communication $\frac{PQ}{2N}$ elements are exchanged. The exchange algorithm routes elements from a node to all other nodes according to an SBT. The SBT's rooted at different nodes are *translations* of each other. A tree rooted at node s is a translation of the tree rooted at node zero, if the addresses of the nodes in the tree rooted at node s are obtained through a bit-wise exclusive-or operation, $x \oplus s$, for every node x of the tree rooted at node zero. In the exchange algorithm the dimensions of the cube can be scanned in an arbitrary order. Starting with the highest-order dimension of the real processor address and virtual processor address before the communication, a single block is communicated in the first transfer. The

number of blocks doubles for each step of the exchange algorithm, and the block size is reduced by a factor of 2.

This exchange algorithm can be explained in terms of the address space of the data set subject to *all-to-all personalized communication.* Let the data assignment before and after the communication be

$$\text{Before}: \quad (\underbrace{w_{m-1}w_{m-2}\cdots w_{rp}}_{vp}\ \underbrace{w_{rp-1}\cdots w_0}_{rp}),$$

$$\text{After}: \quad (\underbrace{w_{m-1}w_{m-2}\cdots w_s}_{vp}\ \underbrace{w_{s-1}w_{s-2}\cdots w_{s-rp}}_{rp}\ \underbrace{w_{s-rp-1}\cdots w_{rp}w_{rp-1}\cdots w_0}_{vp}).$$

Then, in the ith exchange step real processor dimension $rp-i-1$ and virtual processor dimension $s-i-1$, $i \in \{0,1,\cdots,rp-1\}$ are involved in the exchange.

Exchange step i :

$$(\underbrace{w_{m-1}w_{m-2}\cdots w_s}_{vp}\ \underbrace{w_{s-1}\cdots w_{s-i+1}}_{rp}\ w_{s-i}\ \underbrace{w_{s-i-1}\cdots w_{rp-i+1}}_{vp}\ w_{rp-i}\ \underbrace{w_{rp-i-1}\cdots w_0}_{rp}).$$

The data volume in each exchange remains constant, since the number of virtual processor dimensions remain constant. But, the exchange dimension partitions the virtual address space into an increasing number of smaller blocks for increasing i. A shuffle operation on the virtual addresses between each exchange operation would allow the exchange operation to always work with single block exchanges. The shuffle operation implies extensive local data movement.

As an alternative to a local shuffle operation, in order to minimize the number of communication start-ups, blocks can be moved to a buffer, and a number of blocks sent in the same communication. For the Intel iPSC moving data to a buffer requires a significant time, and there exists a block size less than the buffer size for which the copy time is greater than the start-up time. We devised an optimal buffer scheme that is presented in connection with the discussion of our experiments on the Intel iPSC.

DEFINITION 10. The "Standard Exchange Algorithm" on $2l$ dimensions performs an exchange of data between dimensions $g(i)$ and $f(i)$, where the sequences $\{g(i)\}$ and $\{f(i)\}$, $i \in \{0,1,\cdots,l-1\}$, are disjoint and both monotonically increasing, or decreasing, as a function of i. The exchange is made on data such that $w_{g(i)} \oplus w_{f(i)} = 1$.

For instance, $g(i) = s-i-1$ and $f(i) = rp-i-1$ for the above example. If $p = q$, $g(i) = m-1-i$, $f(i) = q-1-i$, and $2l = m$, then the standard exchange algorithm realizes a matrix transposition. There is no particular need to restrict the exchanges to proceed from higher- to lower-order dimensions, or lower- to higher-order dimensions on both virtual and real processor dimensions. By allowing exchanges on arbitrarily paired real and virtual processor dimensions various forms of data conversions can be accomplished. We will give a few examples later (for a general discussion see [4]).

DEFINITION 11. The "General Exchange Algorithm" on $2l$ dimensions performs an exchange between dimensions $g(i)$ and $f(i)$, where $(g(i), f(i))$ is an arbitrary pair of dimensions such that $g(i) \neq g(j)$, $f(i) \neq f(j)$, $i \neq j$, for all $i,j \in \{0,1,\cdots,l-1\}$. An exchange is made on data such that $w_{g(i)} \oplus w_{f(i)} = 1$.

Note that the sets $\{g(i)\}$ and $\{f(i)\}$ are not necessarily disjoint and the sequences $g(0), g(1), \cdots, g(l-1)$ and $f(0), f(1), \cdots, f(l-1)$ are not necessarily increasing or

decreasing. The *general exchange algorithm* can be applied to the *bit-reversal* permutation as described in § 7 and the k shuffle operation described in [4].

With *n-port* communication pipelining can be employed in the exchange algorithm, but the algorithm so modified is suboptimal. However, routing based on spanning balanced n-trees, or rotated spanning binomial trees, and scheduling of data for subtrees in either postorder, or reverse breadth-first order, only requires a time of $T_{\min} = \frac{PQ}{2N} t_c + n\tau$ [14]. A similar algorithm of the same complexity was also derived independently by Stout and Wager [21], [20]. This complexity is again within a factor of 2 of the lower bound

$$T_{l\,b} \geq \max\left(\frac{PQ}{2N} t_c, n\tau\right) \geq \frac{1}{2}\left(\frac{PQ}{2N} t_c + n\tau\right).$$

3.3. All-to-some personalized communication. We only consider the case where $\mathcal{I} = \phi$ and $|\mathcal{R}_b| \neq |\mathcal{R}_a|$. If $|\mathcal{R}_b| = |\mathcal{R}_a| = |\mathcal{R}|$, then the communication is *all-to-all personalized communication.* The general case for which $\mathcal{I} \neq \phi$ is treated in [4]. If the number of real processor dimensions used before the transposition is greater than the number used after the transposition, i.e., if $|\mathcal{R}_b| - |\mathcal{R}_a| = k > 0$, then the transposition implies k steps of *all-to-one personalized communication* and $|\mathcal{R}_a|$ steps of *all-to-all personalized communication.* Data accumulation takes place during the k steps of *all-to-one personalized communication.* If $|\mathcal{R}_a| - |\mathcal{R}_b| = k > 0$, then there are k steps of *one-to-all personalized communication* and $|\mathcal{R}_b|$ steps of *all-to-all personalized communication.* The k steps of *one-to-all personalized communication* imply data splitting.

THEOREM 3.1. *The steps of all-to-one and all-to-all personalized communication used to realize all-to-some personalized communication can be performed in any order. Performing the all-to-all personalized communication first minimizes the data transfer time. For some-to-all personalized communication performing the one-to-all personalized communication first minimizes the data transfer time.*

The theorem simply states that data accumulation shall be performed last and data splitting first. The theorem can be proved by considering the communication complexity of inserting the k steps all-to-one (one-to-all) personalized communication among the all-to-all personalized communication.

Let $k = |(|\mathcal{R}_b| - |\mathcal{R}_a|)|$ and $l = \min(|\mathcal{R}_b|, |\mathcal{R}_a|)$. If the minimized algorithm is executed, for the k steps of *all-to-one* or *one-to-all personalized communication* there are 2^l distinct subcubes in which the operation takes place concurrently. Each such subcube is of dimension k. Also, the *all-to-all personalized communication* takes place within subcubes of dimension l, and there are 2^k such subcubes.

The complexity estimates for $k = |(|\mathcal{R}_b| - |\mathcal{R}_a|)|$ steps of accumulation/splitting and $l = \min(|\mathcal{R}_b|, |\mathcal{R}_a|)$ steps of *all-to-all personalized communication* are given in Table 3.3. Note that $l = n$, $k = 0$ yields the complexity of the *all-to-all personalized communication*, and $l = 0$, $k = n$ yields the complexity of the *one-to-all* or *all-to-one personalized communication.* In general, it is a 2^l-to-2^{l+k} (or 2^{l+k}-to-2^l) personalized communication.

4. Matrix transposition. We have defined matrix transposition as a set of shuffle operations. This definition is convenient on certain processor networks, and for parts of the analysis. Matrix transposition implies an exchange of the row and column address fields. This exchange can clearly be accomplished by the *standard exchange*

TABLE 3
Estimated communication time for some-to-all personalized communication.

Comm. cap.	Time
one-port	$T = (l\frac{PQ}{2^{k+l+1}} + \sum_{i=0}^{k-1} \frac{PQ}{2^{k+l-i}})t_c + (l\lceil \frac{PQ}{B_m 2^{k+l+1}} \rceil + \sum_{i=0}^{k-1} \lceil \frac{PQ}{B_m 2^{k+l-i}} \rceil)\tau$
n-port	$T = (\frac{PQ}{2^{k+l+1}} + \frac{1}{k} \sum_{i=0}^{k-1} \frac{PQ}{2^{k+l-i}})t_c + (l\lceil \frac{PQ}{lB_m 2^{k+l+1}} \rceil + \sum_{i=0}^{k-1} \lceil \frac{PQ}{kB_m 2^{k+l-i}} \rceil)\tau$

algorithm, if $p = q$. If this is not the case, then virtual elements can be introduced to square up the matrix. A standard exchange algorithm can be formulated as follows:

```
For i := q − 1 downto 0
        forall u_i ⊕ v_i = 1
                exchange elements {(u||v)} and {(v||u)};
        endforall
endfor
```

LEMMA 4.1. *Let $p = q$, $u_j = v_j$, for all $j \in \{0, 1, \cdots, i-1, i+1, \cdots, q-1\}$, $u_i = \overline{v_i}$; then* Hamming$((u||v), (v||u)) = 2$.

COROLLARY 4.2. *If the number of processors is equal to the number of matrix elements, 2^m, then matrix transposition performed through a sequence of exchanges requires $\frac{1}{2}m$ exchanges, each requiring communication over a distance of two.*

Corollary 4.2 gives an upper bound equal to the lower bound of Corollary 2.4.

With a one-dimensional partitioning of the matrix, $I = \phi$ regardless of the assignment schemes before and after the transposition. In the two-dimensional partitioning I may be empty, but it can also be equal to the full processor set $\mathcal{R}$.

LEMMA 4.3. *If the exchange algorithm is used for transposition and $g(i), f(i) \in \mathcal{R}_b$, then the communication is between real processors at distance 2. If $g(i) \in \mathcal{R}_b$, $f(i) \notin \mathcal{R}_b$, or $g(i) \notin \mathcal{R}_b$, $f(i) \in \mathcal{R}_b$, then the communication is between real processors at distance 1. Otherwise, the exchange operation is a local data movement.*

5. One-dimensional matrix partitioning. If there are data elements for every real processor both before and after the data rearrangement, then the matrix transposition is *all-to-all personalized communication.* Each node sends $\frac{PQ}{N^2}$ elements to every other node. The communication is *all-to-all personalized communication* regardless of whether or not the scheme for assigning elements to processors is the same before or after the transposition.

If the exchange algorithm is used for *all-to-all personalized communication* then the exchange operations takes place either between a virtual processor and a real processor or two virtual processors in the same real processor. With the same number of processors being used before and after the transposition and *one-port* communication the exchange algorithm is optimum within a factor of 2.

For matrix transposition by the exchange algorithm [9] presented next it is assumed that the matrix is partitioned consecutively by rows and that processor i initially holds the elements of the ith block row. After the transpose operation it will hold the elements of the ith block column. Note that the number of rows in a block row is different from the number of columns in a block column, unless $P = Q$. However, the number of elements in a block row and a block column are the same. For the transpose operation the block row of each processor is partitioned by columns

into N same-sized blocks. The transpose is formed by processor i exchanging its jth block with the ith block of processor j. The data array in each processor holding the elements of a block row is two-dimensional, unless the number of rows is equal to the number of processors, and the local data array after the transposition is also two-dimensional, unless the number of columns is less than or equal to the number of processors. To complete the transposition after the interprocessor communication is completed, this two-dimensional data array can be transposed further locally, explicitly, or implicitly by indirect addressing.

```
/* Transposition by the Standard Exchange Algorithm: */
for j := n − 1 downto 0
    if (bit j of my-addr = 0) then
        exchange blocks ½N to N − 1 of my blocked array
            with my neighbor in dimension j
    else
        exchange blocks 0 to ½N − 1 of my blocked array
            with my neighbor in dimension j
    endif;
    shuffle my blocked array;
endfor
```

The loop can also be performed with the loop index running in the opposite order, but then the first operation in the loop shall be an unshuffle operation, which replaces the shuffle operation at the end of the loop.

For *n-port* communication the exchange algorithm is no longer optimal. An SBnT algorithm as described below yields a communication complexity that is optimum within a factor of 2.

```
/* Transposition by an SBnT Algorithm: */
/* Let the format of msg be (source-addr, relative-addr, data). */
/* base(j) = the minimum number of right rotation of j which yields */
/* the minimum value among all rotations of j. */
for all j ≠ my-addr do
    form msg for processor j = (my-addr, my-addr ⊕ j ⊕ 00···01_b0···0, data)
        and append to output-buf [b], where b is the base of my-addr ⊕ j.
loop n times
    send concurrently for all n output ports.
    receive concurrently for all n input ports.
    for each j do, 0 ≤ j < n
        for each msg of input-buf [j] do
            if relative-addr = 0 then
                put the data into the source-addrth block
                    of the target array
            else
                form relative-addr := relative-addr ⊕ (0···01_p0···0) in
                    the msg and append to output-buf [p], where p is
                    the bit position of relative-addr which is the
                    nearest 1-bit to the left of the jth bit cyclically.
                /* Note: jth bit is always 0 here. */
```

```
        endif
      endfor
    endfor
endloop
```

In the case where only a few processing nodes contain data before or after the transformation it is of the form *some-to-all* or *all-to-some personalized communication.* In the extreme case it is *one-to-all* or *all-to-one personalized communication. Virtual elements* can be introduced such that the same number of real processors are used before and after the transposition. Real processors with virtual elements participate in the exchange operations by receiving data. Virtual elements need not be communicated. The number of real elements being communicated in an exchange operation is not constant, in general, when virtual elements are introduced.

COROLLARY 5.1. *In a one-dimensional partitioning such that* $|\mathcal{R}_b| = |\mathcal{R}_a|$ *there exist elements that must traverse* $|\mathcal{R}_b|$ *dimensions.*

The communication complexity for these cases is summarized in Table 3.3.

LEMMA 5.2. *One-dimensional transposition can be combined with change of data assignment scheme in using the standard exchange algorithm.*

COROLLARY 5.3. *The conversion of the storage form of a matrix stored in* $2^{|\mathcal{R}_b|} \leq 2^n$ *processors from any one of the following storage forms:*

- *consecutive row*
- *consecutive column*
- *cyclic row*
- *cyclic column*
- *combined cyclic and consecutive row storage*
- *combined cyclic and consecutive column storage*

to any other of these forms requires communication from each of the processors to $2^{|\mathcal{R}_a|} - 1$ *other processors, if* $I = \phi$.

COROLLARY 5.4. *The conversion between the cyclic and consecutive storage forms implies all-to-all personalized communication, if* $P \geq N^2$ *for partitioning by rows and* $Q \geq N^2$ *for partitioning by columns.*

For both the SBT and SBnT algorithms presented above it is assumed that the partitions are embedded in the cube by a binary encoding. For Gray code encoding of partitions and binary encoding of virtual processors, we can first perform a transformation locally such that block w is moved to block location $G(w)$, and then carry out the above algorithms. The two operations can also be combined as described in § 6.2.

6. Two-dimensional partitioning. In the two-dimensional partitioning with cyclic assignment and the same number of dimensions for rows and columns the address field is partitioned as follows:

$$(\underbrace{u_{p-1}u_{p-2}\cdots u_{n_c}}_{vp}\ \underbrace{u_{n_c-1}\cdots u_0}_{rp}\ \underbrace{v_{q-1}v_{q-2}\cdots v_{n_c}}_{vp}\ \underbrace{v_{n_c-1}\cdots v_0}_{rp}).$$

For consecutive assignment the partitioning is

$$(\underbrace{u_{p-1}u_{p-2}\cdots u_{p-n_c}}_{rp}\ \underbrace{u_{p-n_c-1}\cdots u_0}_{vp}\ \underbrace{v_{q-1}v_{q-2}\cdots v_{q-n_c}}_{rp}\ \underbrace{v_{q-n_c-1}\cdots v_0}_{vp}).$$

In either of these cases $I = \mathcal{R}_b = \mathcal{R}_a$. This case is the basic two-dimensional matrix transposition. The communication is between pairs of processors. In [8], [9]

we show that the transposition of a matrix embedded in the cube by a binary code or Gray code encoding implies the same communication. The algorithm in the above references uses a single path from source to destination for every source/destination pair. We will describe a simple extension using two paths for every source/destination pair, and an algorithm using multiple paths. We refer to the three algorithms by the names *Single Path Transpose* (SPT), *Dual Path Transpose* (DPT), and *Multiple Path Transpose* (MPT). Note that by Corollary 2.4 the maximum distance matrix elements need to traverse is $n = 2n_c$.

With a mixed assignment before and after the transposition, such as consecutive for rows and cyclic for columns

$$(\underbrace{u_{p-1}u_{p-2}\cdots u_{p-n_r}}_{rp}\ \underbrace{u_{p-n_r-1}\cdots u_0}_{vp}\ \underbrace{v_{q-1}v_{q-2}\cdots v_{n_c}}_{vp}\ \underbrace{v_{n_c-1}\cdots v_0}_{rp}),$$

the set $\mathcal{I}$ may no longer equal the entire processor set. In fact, if $q - n_c \geq n_r$ and $p - n_r \geq n_c$ then $\mathcal{I} = \phi$ and it is an *all-to-all personalized communication.* Between these two extremes the communication is discussed in [4].

6.1. Transposition with communication only between distinct source/destination pairs of processors. We consider the transpose operation for binary encoding first. Define $\mathrm{tr}(x)$ to be the function which maps the address of partition (address of assigned processor) $x = (x_r||x_c)$ to the address of the transposed partition, i.e., $\mathrm{tr}(x) = (x_c||x_r)$. Let $H(x) = \mathrm{Hamming}(x_r, x_c)$. Then $\mathrm{Hamming}(x, \mathrm{tr}(x)) = 2H(x)$. In the following we assume that $n_r = n_c = \frac{n}{2}$ and n is even, i.e., that there are equally many row and column partitions.

The *Single Path Transpose* (SPT) algorithm [9], [15] uses one path from processor x to processor $\mathrm{tr}(x)$. Paths for different $x's$ are edge-disjoint, and pipelining of communications can be employed to reduce the communication complexity. The *Dual Paths Transpose* (DPT) algorithm is a straightforward improvement of the SPT algorithm in that two directed edge-disjoint paths are established from each source processor to its corresponding destination processor. In the *Multiple Paths Transpose* (MPT) algorithm, we partition the processor addresses into sets such that all members of a set have equivalent properties with respect to an relation operator (defined later). We show that the paths associated with any two source processors belonging to different sets are edge-disjoint. We then prove that all the paths of the processors in the same set share the same set of edges, but we use them during different cycles. An algorithm similar to the MPT algorithm was also derived independently by Stout and Wager [20], [21].

6.1.1. The Single Path Transpose (SPT) algorithm. With the same assignment scheme for rows and columns, and the same number of processors assigned to rows and columns, $n_c = n_r = \frac{n}{2}$ (n must be even) the communication is restricted to distinct source/destination pairs. The *Single Path Transpose* (SPT) algorithm [9], [15] is a special case of the standard exchange algorithm.

LEMMA 6.1. *In a two-dimensional partitioning such that the same number of dimensions are used for real processor addresses before and after the transposition and the same assignment scheme used before and after the transposition there exist elements that must traverse $rp = 2n_c$ dimensions.*

In the SPT algorithm for the same number of real processors for rows and columns, and the same assignment scheme for both rows and columns both before and after the transposition, data is exchanged between processors with addresses that differ in

dimensions $g(i)$, $g(i) \in \mathcal{R}, g(i) < q$, and $f(i)$, $f(i) \in \mathcal{R}, q \leq f(i) < m$ in the ith exchange step. The implied routing corresponds to directed edge-disjoint paths from each node x to $\mathrm{tr}(x)$. For each source-destination pair there is a single path. This path only goes through the appropriate dimensions of the real processor addresses corresponding to the bits of x that need to be complemented to become the destination real address $\mathrm{tr}(x)$. The routing order for the dimensions is the same for all nodes, for instance, highest- to lowest-order for both row and column encoding, i.e., $g(\frac{n}{2}-1)$, $f(\frac{n}{2}-1)$, $g(\frac{n}{2}-2)$, $f(\frac{n}{2}-2), \cdots, g(0)$, $f(0)$. The length of the path of node x is $2H(x)$. The first packet for each node on the antidiagonal arrives after n routing steps and additional packets every cycle thereafter. The total number of routing steps is $\lceil \frac{PQ}{BN} \rceil + n - 1$. The nodes which are not on the antidiagonal can either finish the transposition earlier in a "greedy" manner, or synchronize with the antidiagonal nodes, i.e., the packet with the same ordinal number of all the nodes uses the same dimension (or idles) during the same step. The total transposition time T is $(\lceil \frac{PQ}{BN} \rceil + n-1)(Bt_c+\tau)$. The optimal packet size B_{opt} is $\sqrt{PQ\tau/N(n-1)t_c}$ and the minimum time is $T_{\min} = (\sqrt{\frac{PQ}{N}t_c} + \sqrt{(n-1)\tau})^2$.

6.1.2. The Dual Paths Transpose (DPT) algorithm. The SPT algorithm can be improved by establishing two directed edge-disjoint paths between x and $\mathrm{tr}(x)$ for all x's. In addition to the paths used in the SPT algorithm, a second path is defined by permuting processor row and column dimensions pairwise to yield a routing order selected from $f(\frac{n}{2}-1)$, $g(\frac{n}{2}-1)$, $f(\frac{n}{2}-2)$, $g(\frac{n}{2}-2), \cdots, f(0)$, $g(0)$. The two directed paths for a particular x are edge-disjoint (as observed in [10] for the solution of tridiagonal systems on Boolean cubes). Moreover, the two directed paths for any x are edge-disjoint with respect to all paths for other x's. This second path can be used to reduce the time for data transfer by splitting the set of data $\frac{PQ}{N}$ into two equal parts. The path lengths are already minimal in the SPT algorithm. The communication complexity is $(\lceil \frac{PQ}{2BN} \rceil + n-1)(Bt_c+\tau)$, which is minimized for $B = B_{\mathrm{opt}} = \sqrt{\frac{PQ\tau}{2N(n-1)t_c}}$ and $T_{\min} = (\sqrt{\frac{PQ}{2N}t_c} + \sqrt{(n-1)\tau})^2$. The speed-up is approximately 2 for $\frac{PQ}{N}t_c \gg n\tau$, i.e., for Boolean cubes small relative to the problem size. Note that for the SPT algorithm it suffices that each node supports a total of n concurrent send or receive operations, whereas for the DPT algorithm n send operations concurrently with n receive operations are required for each node. Unidirectional communication suffices for the SPT algorithm, but bidirectional communication is required for the DPT algorithm.

6.1.3. The Multiple Paths Transpose (MPT) algorithm. For the *Multiple Paths Transpose* (MPT) algorithm we define $2H(x)$ paths, labeled $0, 1, \cdots, 2H(x)-1$, between nodes x and $\mathrm{tr}(x)$. The paths differ in the order in which the dimensions are routed. All paths originated from the same node have the same length. Let $\alpha_{H(x)-1}$, $\alpha_{H(x)-2}$, $\cdots$, α_0, $\beta_{H(x)-1}$, $\beta_{H(x)-2}$, $\cdots$, β_0 be the sequence of dimensions that need to be routed in descending order. We describe a path as a sequence of dimensions:

$$\text{path } p = \begin{cases} \alpha_{(p+H(x)-1) \bmod H(x)}, \beta_{(p+H(x)-1) \bmod H(x)}, \alpha_{(p+H(x)-2) \bmod H(x)}, \beta_{(p+H(x)-2) \bmod H(x)}, \\ \qquad \cdots, \alpha_p, \beta_p, \quad \forall p \in \{0, 1, \cdots, H(x)-1\}, \\ \beta_{(j+H(x)-1) \bmod H(x)}, \alpha_{(j+H(x)-1) \bmod H(x)}, \beta_{(j+H(x)-2) \bmod H(x)}, \alpha_{(j+H(x)-2) \bmod H(x)}, \\ \qquad \cdots, \beta_j, \alpha_j, \quad j = p - H(x) \quad \forall p \in \{H(x), H(x)+1, \cdots, 2H(x)-1\}. \end{cases}$$

For example, if $x = (1001||0100)$, then $x_r = 1001, x_c = 0100, H(x) = 3$ and $\mathrm{tr}(x) = (x_c||x_r) = (0100||1001)$. The distance between x and $\mathrm{tr}(x)$ is 6. The six paths are defined as follows:

$$\begin{array}{ll} \text{path } 0 = 7,3,6,2,4,0 & \text{path } 3 = 3,7,2,6,0,4 \\ \text{path } 1 = 4,0,7,3,6,2 & \text{path } 4 = 0,4,3,7,2,6 \\ \text{path } 2 = 6,2,4,0,7,3 & \text{path } 5 = 2,6,0,4,3,7. \end{array}$$

Path 0 starts from the source node (10010100) and goes through nodes (00010100), (00011100), (01011100), (01011000), (01001000) and reaches the destination node (01001001). Path p can be derived by a right rotation of two steps of path $(p-1) \bmod H(x)$, if $0 \le p < H(x)$. For $H(x) \le p < 2H(x)$, path p can be derived by a right rotation of two steps of path $((p-1) \bmod H(x)) + H(x)$ and also by permuting row and column dimensions pairwise of path $p \bmod H(x)$. Note that path 0 is the same as the path defined in the SPT algorithm. Paths 0 and $H(x)$ are the two paths defined for node x in the DPT algorithm.

DEFINITION 12. Let x', x'' be two nodes with $x' = (x'_r||x'_c)$ and $x'' = (x''_r||x''_c)$. Define a relation $\sim_{\mathrm{ad}}$ between x' and x'' such that $x' \sim_{\mathrm{ad}} x''$ if and only if $x'_r + x'_c = x''_r + x''_c$, i.e., x' and x'' are on the same antidiagonal. Note that if $x' \sim_{\mathrm{ad}} x''$ and $x'' \sim_{\mathrm{ad}} x'''$ then $x' \sim_{\mathrm{ad}} x'''$.

DEFINITION 13. Define $edge(x,p,e)$ to be the function which returns the eth directed edge of path p of node x, with $e \ge 1$. We also define Edges, OddEdges, EvenEdges and Paths as follows:

$$\begin{aligned} &\mathrm{Edges}(x,e) = \{edge(x,p,e) | \forall p \in \{0,1,\cdots,2H(x)-1\}\}, \\ &\mathrm{OddEdges}(x) = \bigcup_{\forall \text{ odd } e} \mathrm{Edges}(x,e), \\ &\mathrm{EvenEdges}(x) = \bigcup_{\forall \text{ even } e} \mathrm{Edges}(x,e), \\ &\mathrm{Paths}(x) = \mathrm{OddEdges}(x) \bigcup \mathrm{EvenEdges}(x). \end{aligned}$$

DEFINITION 14. Define $\mathrm{Nodes}(x,e)$ to be the function which returns the set of nodes upon which the directed edges in $\mathrm{Edges}(x,e)$ terminate. Define $\mathrm{OddNodes}(x)$ and $\mathrm{EvenNodes}(x)$ to be the set of nodes on which the set of directed edges $\mathrm{OddEdges}(x)$ and $\mathrm{EvenEdges}(x)$ terminate, respectively.

$$\begin{aligned} \mathrm{OddNodes}(x) &= \bigcup_{\forall \text{ odd } e} \mathrm{Nodes}(x,e), \\ \mathrm{EvenNodes}(x) &= \bigcup_{\forall \text{ even } e} \mathrm{Nodes}(x,e). \end{aligned}$$

DEFINITION 15. Let x', x'' be two nodes. Define a relation $\sim_s$ such that $x' \sim_s x''$ if and only if $x' \sim_{\mathrm{ad}} x''$ and $x' \oplus \mathrm{tr}(x') = x'' \oplus \mathrm{tr}(x'')$. Note that if $x' \sim_s x''$ and $x'' \sim_s x'''$ then $x' \sim_s x'''$.

$x' \oplus \mathrm{tr}(x') = x'' \oplus \mathrm{tr}(x'')$ implies $H(x') = H(x'')$, but $H(x') = H(x'')$ does not imply $x' \oplus \mathrm{tr}(x') = x'' \oplus \mathrm{tr}(x'')$. There exists x', x'' such that $x' \sim_{\mathrm{ad}} x''$ and $x' \oplus \mathrm{tr}(x') \ne x'' \oplus \mathrm{tr}(x'')$, for instance (001||111) and (010||110). Also there exists

x', x'' such that $x' \not\sim_{\text{ad}} x''$ and $x' \oplus \text{tr}(x') = x'' \oplus \text{tr}(x'')$, for instance (001||111) and (000||110).

DEFINITION 16. A set of paths defined upon a set of nodes $\mathcal{X}$ is said to be (t, n)-disjoint, $t \leq n$, if a packet that can be transmitted in unit time can be sent out on every path from every node $x \in \mathcal{X}$ during cycles $i * n + 1$, $i * n + 2$, $\cdots$, $i * n + t$, for all $i \geq 0$, without routing conflicts, i.e., messages originating from different nodes will not be routed over the same edge during the same cycle.

Note that the (t, n)-disjoint definition does not imply that the paths from the different source nodes are edge-disjoint, unless $t = n$.

To describe the MPT algorithm we first prove the following properties.

1. Paths p_1 and p_2 of node x are edge-disjoint, for all $p_1, p_2 \in \{0, 1, \cdots, 2H(x)-1\}$, $p_1 \neq p_2$.
2. If $x' \not\sim_s x''$ then $\text{Paths}(x') \cap \text{Paths}(x'') = \phi$.
3. The set of all paths for the nodes in the set induced by the relation $\sim_s$ is $(2, 2H(x))$-disjoint, where x is in the node set.

LEMMA 6.2. *Paths p_1 and p_2 of node x are edge-disjoint, for all $p_1, p_2 \in \{0, 1, \cdots, 2H(x) - 1\}$, $p_1 \neq p_2$.*

Proof. It follows from the facts that all the paths are pointing away from the source node and no two paths traverse the same dimension during the same step. □

LEMMA 6.3. *If $H(x') > 0$, then the set of nodes* $\text{OddNodes}(x')$ *and* $\text{EvenNodes}(x')$ *have the following properties:*

- $x' \not\sim_{\text{ad}} x''$, $H(x'') = H(x') - 1$, *for all* $x'' \in \text{OddNodes}(x')$;
- $x' \sim_{\text{ad}} x''$, $x' \oplus \text{tr}(x') = x'' \oplus \text{tr}(x'')$ (*which implies* $H(x'') = H(x')$), *for all* $x'' \in \text{EvenNodes}(x')$.

Proof. In traversing an edge in $\text{OddEdges}(x)$, we complement one of the $H(x)$ bits of the $\frac{n}{2}$ high- (low-) order bits, which differ from the corresponding low- (high-) order bit. In traversing an edge in $\text{EvenEdges}(x)$, we complement the low- (high-) order bit of the corresponding high- (low-) order bit that was complemented in traversing the preceding odd edge. Let x', x'', and x''' be nodes along the same path such that $x' = (y'||z') \in \text{Nodes}(x, 2h)$, $x'' = (y''||z'') \in \text{Nodes}(x, 2h + 1)$ and $x''' = (y'''||z''') \in \text{Nodes}(x, 2h + 2)$, for all $h \in \{0, 1, \cdots, H(x) - 1\}$. From the definition of paths either $y'' = y' + 2^k$, $z'' = z'$ or $y'' = y'$, $z'' = z' - 2^k$ for some k satisfying $y'_k = 0$, $z'_k = 1$; or $y'' = y' - 2^k$, $z'' = z'$ or $y'' = y'$, $z'' = z' + 2^k$ for some k satisfying $y'_k = 1$, $z'_k = 0$. These conditions imply $y' + z' \neq y'' + z''$, i.e., $x' \not\sim_{\text{ad}} x''$, and $\text{Hamming}(y'', z'') = \text{Hamming}(y', z') - 1$, i.e., $H(x'') = H(x') - 1$. Furthermore, $y''' = y' + 2^k$, $z''' = z' - 2^k$ for some k satisfying $y'_k = 0$, $z'_k = 1$ or $y''' = y' - 2^k$, $z''' = z' + 2^k$, for some k satisfying $y'_k = 1$, $z'_k = 0$. Hence, $y' + z' = y''' + z'''$, i.e., $x' \sim_{\text{ad}} x'''$. Also, $y' \oplus z' = y''' \oplus z'''$, i.e., $(y'||z') \oplus (z'||y') = (y'''||z''') \oplus (z'''||y''')$, which implies $x' \oplus \text{tr}(x') = x''' \oplus \text{tr}(x''')$. □

COROLLARY 6.4. $x' \sim_s x''$, *for all* $x'' \in \text{EvenNodes}(x')$.

LEMMA 6.5. *If $x' \not\sim_{\text{ad}} x''$, then* $\text{Paths}(x') \cap \text{Paths}(x'') = \phi$.

Proof. It is sufficient to prove $\text{Paths}(x') \cap \text{Paths}(x'') = \phi$ by proving $\text{EvenNodes}(x') \cap \text{EvenNodes}(x'') = \phi$ and $\text{EvenNodes}(x') \cap \text{OddNodes}(x'') = \phi$. From Lemma 6.3, $\text{EvenNodes}(x') \sim_{\text{ad}} x'$, $\text{EvenNodes}(x'') \sim_{\text{ad}} x''$. Since $x' \not\sim_{\text{ad}} x''$, we have $\text{EvenNodes}(x') \not\sim_{\text{ad}} \text{EvenNodes}(x'')$, which implies $\text{EvenNodes}(x') \cap \text{EvenNodes}(x'') = \phi$.

To prove $\text{EvenNodes}(x') \cap \text{OddNodes}(x'') = \phi$, we consider three cases.

1. If $H(x') = H(x'')$, then, by Lemma 6.3, $H(y') = H(y'') + 1$, where $y' \in \text{EvenNodes}(x')$, $y'' \in \text{OddNodes}(x'')$. Therefore, $\text{EvenNodes}(x') \cap$

$\text{OddNodes}(x'') = \phi$.

2. If $H(x') > H(x'')$, then $H(y') > H(y'')$, where $y' \in \text{EvenNodes}(x')$, $y'' \in \text{OddNodes}(x'')$. Therefore, $\text{EvenNodes}(x') \cap \text{OddNodes}(x'') = \phi$.
3. If $H(x') < H(x'')$, we show that $\text{EvenNodes}(x'') \cap \text{OddNodes}(x') = \phi$ instead by a similar argument as in case 2. □

LEMMA 6.6. *If* $x' \sim_{\text{ad}} x''$ *and* $x' \not\sim_s x''$, *then* $\text{Paths}(x') \cap \text{Paths}(x'') = \phi$.

Proof. Assume $\text{EvenNodes}(x') \cap \text{EvenNodes}(x'') \neq \phi$. Then there exists one node y such that $y \in \text{EvenNodes}(x')$ and $y \in \text{EvenNodes}(x'')$. By Corollary 6.4, $y \sim_s x', y \sim_s x''$, i.e., $x' \sim_s x''$ which is a contradiction. So, $\text{EvenNodes}(x') \cap \text{EvenNodes}(x'') = \phi$. Also by Lemma 6.3, $y' \not\sim_{\text{ad}} y''$, for all $y' \in \text{EvenNodes}(x')$ and $y'' \in \text{OddNodes}(x'')$, which means $\text{EvenNodes}(x') \cap \text{OddNodes}(x'') = \phi$. Hence, $\text{Paths}(x') \cap \text{Paths}(x'') = \phi$. □

LEMMA 6.7. *If* $x' \not\sim_s x''$ *then* $\text{Paths}(x') \cap \text{Paths}(x'') = \phi$.

Proof. The proof follows from Lemmas 6.5 and 6.6. □

LEMMA 6.8. *The set of paths defined for the nodes in the same set induced by the relation* $\sim_s$ *is* $(2, 2H(x))$*-disjoint.*

Proof. We first prove that the paths of the nodes defined by the relation $\sim_s$ are $(1, 2H(x))$-disjoint. The proof is by induction on the routing cycles. During cycles 1 and 2, the routed edges are clearly disjoint by Lemma 6.3. Assume that during cycles $2n-1$ and $2n$, $n > 0$, the routing is also edge-disjoint. If $n = H(x)$, then all the routing is complete. During the next two cycles the routing is restarted and there is no edge conflict. If $n \neq H(x)$, then consider the $2H(x)$ edges directed into some node y at distance $2n$ from x as well as the $2H(x)$ edges directed out from node y. Let $\alpha_{H(x)-1}, \alpha_{H(x)-2}, \cdots, \alpha_0, \beta_{H(x)-1}, \beta_{H(x)-2}, \cdots, \beta_0$ be the corresponding $2H(x)$ dimensions in descending order. If an edge used during cycle $2n-1$ is in dimension α_k (i.e., the edge used during cycle $2n$ is in dimension β_k), then the edges used during the following two cycles are in dimensions $\alpha_{(k-1) \bmod H(x)}$ and $\beta_{(k-1) \bmod H(x)}$, respectively. If the edge used during cycle $2n-1$ is in dimension β_k, then the edges used during the following two cycles are in dimensions $\beta_{(k-1) \bmod H(x)}$ and $\alpha_{(k-1) \bmod H(x)}$, respectively. Hence, the edges used during the following two cycles are all distinct and it follows that the paths are $(1, 2H(x))$-disjoint.

To show that the paths are $(2, 2H(x))$-disjoint it suffices to show that the set of edges used during odd cycles (odd edges) are disjoint from the set of edges used during even cycles (even edges). Let x be any node in the set defined by the relation $\sim_s$. That the set of edges used during odd cycles are disjoint from the set of edges used during even cycles follows from the property that odd edges are directed from node x' to node y' and even edges are directed from node y'' to node x'', where $x \sim_s x' \sim_s x''$, $x \not\sim_s y'$ and $x \not\sim_s y''$. □

Figure 3 shows an instance of a set induced by the relation $\sim_s$ on a 6-cube. Note that $H(x) = 3$ for x in this set. The nodes in the same set form a *logical* $H(x)$-dimensional cube, where each logical link represents an exchange operation of two dimensions. Hence, a logical link contains two disjoint paths of length 2. By Lemma 6.7, the corresponding physical edges of the logical link will only be shared by nodes in this set. Notice that x and $\text{tr}(x)$ are at maximum distance from each other in the logical $H(x)$-cube. Figure 4 shows the six $(2H(x))$ edge-disjoint paths from node $x = (000111)$ to node $\text{tr}(x) = (111000)$. The labels on the edges are dimensions of the edges.

For the routing, the data from node x is split into $4H(x)$ packets of size $\lceil \frac{PQ}{4NH(x)} \rceil$ each. The packets are sent during the first two cycles. The first $2H(x)$ packets will

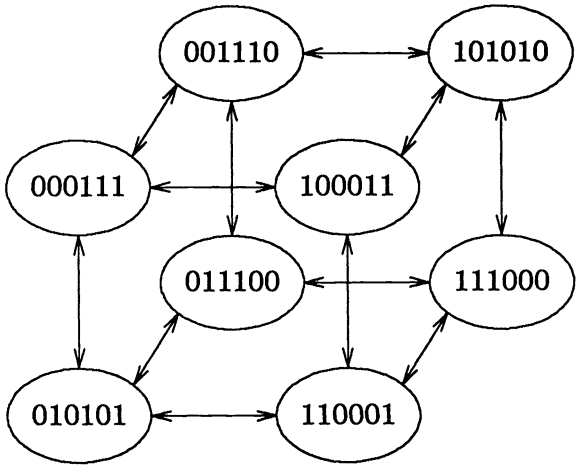


FIG. 3. *The logical $H(x)$-cube formed by nodes in the same induced set of the relation $\sim_s$.*

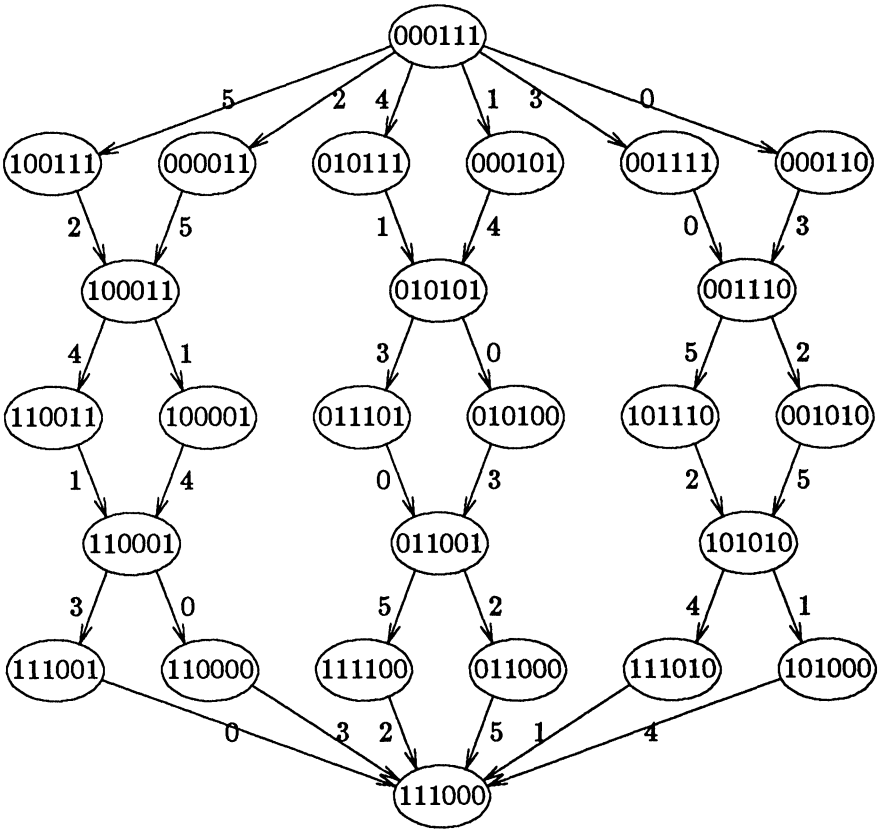


FIG. 4. *Six $(2H(x))$ edge-disjoint paths from node $x = (000111)$ to node $\mathrm{tr}(x) = (111000)$.*

arrive at the destination node, $\mathrm{tr}(x)$, after $2H(x)$ cycles, and the second set during the next cycle. The total transpose time is

$$\begin{cases} (n+1)\tau + (\frac{n+1}{2n})\frac{PQ}{N}t_c & \text{if } \frac{n}{2} \geq \frac{PQt_c}{8N\tau}, \\ 3\tau + \frac{3}{4}\frac{PQ}{N}t_c & \text{otherwise.} \end{cases}$$

The transpose time decreases as a function of $H(x)$ for $1 \leq H(x) \leq \sqrt{\frac{PQt_c}{8N\tau}}$ and increases for $\sqrt{\frac{PQt_c}{8N\tau}} \leq H(x)$. The transpose time for $H(x) = 1$ and $H(x) = \frac{PQt_c}{8N\tau}$ are the same. The maximal packet size is $\frac{PQ}{4N}$. The maximal packet size can be reduced by splitting the data into $\lfloor \frac{n}{2H(x)} \rfloor * 4H(x)$ packets. The total transpose time either remains unchanged (if $\frac{n}{2} \geq \frac{PQt_c}{8N\tau}$) or is reduced. In fact, the data sent from node x can be split into $4kH(x)$ packets instead of $4H(x)$ packets. The whole routing completes in $2kH(x)+1$ cycles. Hence, $T = (2kH(x)+1)(\tau + \frac{PQt_c}{4kH(x)N})$, $H(x) \in \{1, 2, \cdots, \frac{1}{2}n\}$. The optimal k is $\frac{1}{2H(x)}\sqrt{\frac{PQt_c}{2N\tau}}$ and $T_{\min} = (\sqrt{\tau} + \sqrt{\frac{PQt_c}{2N}})^2$. Notice that $T_{\min}$ is valid only when $k \geq 1$, which implies $\sqrt{\frac{PQt_c}{2N\tau}} \geq n$.

THEOREM 6.9. *The total matrix transpose time by the MPT algorithm is*

$$T_{\min} = \begin{cases} (n+1)\tau + \frac{n+1}{2n}\frac{PQ}{N}t_c & \textit{if } n \geq \sqrt{\frac{PQt_c}{N\tau}} \textit{ approximately,} \\ (\frac{n}{2}+3)\tau + \frac{n+6}{2n+8}\frac{PQ}{N}t_c & \textit{if } \sqrt{\frac{PQt_c}{2N\tau}} < n \leq \sqrt{\frac{PQt_c}{N\tau}} \textit{ approximately and } \frac{n}{2} \textit{ is even,} \\ (\frac{n}{2}+2)\tau + \frac{n+4}{2n+4}\frac{PQ}{N}t_c & \textit{if } \sqrt{\frac{PQt_c}{2N\tau}} < n \leq \sqrt{\frac{PQt_c}{N\tau}} \textit{ approximately and } \frac{n}{2} \textit{ is odd,} \\ (\sqrt{\tau} + \sqrt{\frac{PQt_c}{2N}})^2 & \textit{if } n \leq \sqrt{\frac{PQt_c}{2N\tau}}, \end{cases}$$

and the optimum packet size is

$$B_{\mathrm{opt}} = \begin{cases} \lceil \frac{PQ}{N(n+4)} \rceil & \textit{for even } \frac{n}{2} \textit{ and } n > \sqrt{\frac{PQt_c}{2N\tau}}, \\ \lceil \frac{PQ}{N(n+2)} \rceil & \textit{for odd } \frac{n}{2} \textit{ and } n > \sqrt{\frac{PQt_c}{2N\tau}}, \\ \sqrt{\frac{PQ\tau}{2Nt_c}} & \textit{for } n \leq \sqrt{\frac{PQt_c}{2N\tau}}. \end{cases}$$

THEOREM 6.10. *The matrix transposition time is at least* $\max(n\tau, \frac{PQ}{2N}t_c)$.

Proof. The minimum number of start-ups is determined by the longest distance, which is n. Nodes on the main antidiagonal are at distance n. For a lower bound on the required time for data transfer consider the upper right $\frac{\sqrt{N}}{2} \times \frac{\sqrt{N}}{2}$ submatrix. There are $\frac{N}{4}$ nodes. Each node has to send $\frac{PQ}{N}$ data to some node outside the submatrix. There are two links per node that connect to nodes outside of the submatrix, i.e., a total of $\frac{2N}{4}$ links. Hence, the data transfer requires a time of at least $\frac{PQ}{2N}t_c$. □

For Gray code encoding on both row and column indices, we can apply exactly the same transpose algorithm. For a binary encoding of row and column indices, matrix element (u, v) is stored in processor $w = (u||v)$ and matrix element (v, u) is stored in processor $\mathrm{tr}(w) = (v||u)$. For Gray code encoding of row and column indices, matrix element (u, v) is stored in processor $(G(u)||G(v))$ and matrix element (v, u) is stored in processor $(G(v)||G(u))$. The two-dimensional transpose algorithms described above are indeed permutation algorithms defined by $(u||v) \leftarrow (v||u)$, for all $u \in \{0, 1, \cdots, P-1\}$, $v \in \{0, 1, \cdots, Q-1\}$. It follows that the permutation will

transpose the matrix. In general, if row and column indices are encoded in the same way, the transpose algorithm only depends on the processor addresses, not on the row and column indices of the matrix elements in the processors. For $N < PQ$, the argument applies to matrix blocks instead of matrix elements.

6.2. Transposition with change of assignment scheme. If the number of processors in the row and column direction are not the same, or if a different assignment strategy is used for rows and columns, or if the assignment scheme after the transpose is different from that before the transpose, then the communication is no longer confined to distinct pairs. If $|\mathcal{R}_b| = |\mathcal{R}_a| = |\mathcal{R}|$ and $I = \phi$, then the communication is *all-to-all personalized communication.* In general, for $I \neq \phi$ the transposition/rearrangement is composed of different types of operations. This case is treated further in [4].

For a nonsquare matrix virtual elements can be introduced. Virtual elements need not be communicated, and the complexity of the transposition is reduced accordingly, but the basic algorithms apply.

To illustrate a two-dimensional transposition with change of assignment scheme, such that $I = \phi$, we consider the transposition of a matrix stored consecutively with respect to both rows and columns before the transposition, and stored cyclically with respect to both rows and columns after the transposition. We also assume that $n_r = n_c$ and that $p, q \geq 2n_r$. The partitioning of the address field before and after the transposition and change of assignment scheme are

$$\text{Before}: \quad (\underbrace{u_{p-1}u_{p-2}\cdots u_{p-n_c}}_{rp}\ \underbrace{u_{p-n_c-1}\cdots u_0}_{vp}\ \underbrace{v_{q-1}v_{q-2}\cdots v_{q-n_c}}_{rp}\ \underbrace{v_{q-n_c-1}\cdots v_0}_{vp}),$$

$$\text{After}: \quad (\underbrace{v_{q-1}v_{q-2}\cdots v_{n_c}}_{vp}\ \underbrace{v_{n_c-1}\cdots v_0}_{rp}\ \underbrace{u_{p-1}u_{p-2}\cdots u_{n_c}}_{vp}\ \underbrace{u_{n_c-1}\cdots u_0}_{rp}).$$

We consider three exchange algorithms that differ only in the way dimensions are paired, and the order in which the exchanges are performed. Let *exchange-row*(M, s, N_r) denote the sequence of exchange operations between N_r block rows (within a column subcube of N_r processors) as defined by the standard exchange algorithm described in pseudocode before, except for a minor modification. The initial local array of length M is partitioned into $2^s N_r$ blocks. The jth block is sent to processor $j \bmod N_r$, for all $j \in \{0, 1, \cdots, 2^s N_r - 1\}$ during the execution of the exchange algorithm. Each processor sends 2^s blocks to every other processor. For the exchange algorithm for the transposition of a one-dimensionally partitioned matrix described earlier, $M = \frac{PQ}{N}$, $s = 0$, $N_r = N$. Each processor sends only one block to every other processor. *Exchange-row*(M, s, N_r) operates within each column subcube. *Exchange-column*(M, s, N_c) is defined analogously.

The parameter s defines the offset from the high-order dimension of the virtual processor address field for the first exchange in the *standard exchange algorithm.* From the discussion of the standard exchange algorithm it is clear that an offset of s divides the local array into 2^{s+1} blocks for the first exchange. The blocks are of size $\frac{PQ}{2^{s+1}N}$ for a $P \times Q$ matrix partioned evenly among N real processors.

For the transposition with change of assignment scheme we consider the following three algorithms:

1. Convert from consecutive-row partitioning to cyclic-row partitioning, i.e., *exchange-row*$(\frac{PQ}{N}, p - 2n_r, N_r)$; then convert from consecutive-column partitioning to cyclic-column partitioning, by employing *exchange-column*$(\frac{PQ}{N}, m-$

$n - n_c, N_c$); then transpose the matrix globally and locally.

2. Transpose the local matrices concurrently; then convert from consecutive-row partitioning to cyclic-row partitioning, i.e., *exchange-row*$(\frac{PQ}{N}, p - 2n_r, N_r)$; then convert from consecutive-column partitioning to cyclic-column partitioning, i.e., *exchange-column*$(\frac{PQ}{N}, m - n - n_c, N_c)$; then transpose N local matrices each of size $2^{p-2n_r} \times 2^{q-2n_c}$ concurrently in all N real processors.
3. Convert from consecutive-column to cyclic-column partitioning between rows (within each column subcube), i.e., *exchange-row*$(\frac{PQ}{N}, m - n - n_c, N_r)$; then convert from consecutive-row to cyclic-row partitioning between columns (within each row subcube), i.e., *exchange-column*$(\frac{PQ}{N}, p - n, N_c)$. A local $p - 2n_r$ shuffle operation is necessary if $p > 2n_r$.

The algorithms can be illustrated in terms of operations on the address field. For simplicity let it be partitioned as $(u_1u_2u_3v_1v_2v_3)$, where u_1, u_3, v_1, and v_3 all define subfields of n_r dimensions. u_1 and v_1 are the real processor address fields before the transposition, u_3 and v_3 the real fields after transposition and change of assignment scheme.

ALGORITHM 1:

$$(\underline{u_1}u_2u_3\underline{v_1}v_2v_3) \rightarrow (u_1u_2\underline{u_3}\underline{v_1}v_2v_3) \rightarrow (u_1u_2\underline{u_3}v_1v_2\underline{v_3}) \rightarrow (v_1v_2\underline{v_3}u_1u_2\underline{u_3}).$$

ALGORITHM 2:

$$(\underline{u_1}u_2u_3\underline{v_1}v_2v_3) \rightarrow (\underline{u_1}v_2v_3\underline{v_1}u_2u_3) \rightarrow (u_1v_2\underline{v_3}\underline{v_1}u_2u_3) \rightarrow$$
$$(u_1v_2\underline{v_3}v_1u_2\underline{u_3}) \rightarrow (v_1v_2\underline{v_3}u_1u_2\underline{u_3}).$$

ALGORITHM 3:

$$(\underline{u_1}u_2u_3\underline{v_1}v_2v_3) \rightarrow (\underline{v_3}u_2u_3\underline{v_1}v_2u_1) \rightarrow (\underline{v_3}u_2v_1\underline{u_3}v_2u_1) \rightarrow (\underline{v_3}v_1v_2\underline{u_3}u_1u_2).$$

The underline denotes the real processor address field. Note that the last form in Algorithm 3, $(\underline{v_3}v_1v_2\underline{u_3}u_1u_2)$, denotes the same assignment scheme as $(v_1v_2\underline{v_3}u_1u_2\underline{u_3})$. The steps of the three different algorithms are illustrated in Fig. 5 in terms of the matrix. The number in the figure denotes (row-index||column-index).

The first algorithm requires $2n$ communication steps, the second only n steps. However, the second algorithm requires a complete local matrix transpose before the interprocessor communication phase, and the transposition of a number of smaller matrices after the communication. The third algorithm also requires n communication steps, but no transposition is required prior to the communication. A local $p - 2n_r$ shuffle operation is required if $p > 2n_r$. Note that the order between the *exchange-row* and *exchange-column* operations can be reversed.

Conversion between cyclic and consecutive assignment in the row or column direction is equivalent to a number (N_c or N_r) of independent one-dimensional conversions. Conversion in both dimensions is equivalent to *all-to-all personalized communication* if $Q \geq N_c^2$ and $P \geq N_r^2$.

6.3. Combining transpose and Gray code/binary code conversion. For the transpose of a matrix with the row index encoded in binary code and the column index in Gray code, a binary-to-Gray-code conversion can first be done for each column subcube concurrently in $\frac{n}{2} - 1$ steps [9], then the Gray-to-binary-code conversion for each row subcube concurrently in another $\frac{n}{2} - 1$ steps followed by the n-step transpose algorithm. The two conversions and the transposition commute. The total number of routing steps is $2n - 2$. However, the number of routing steps can be reduced to n, if the SPT algorithm is used for the transposition by combining it with the conversion

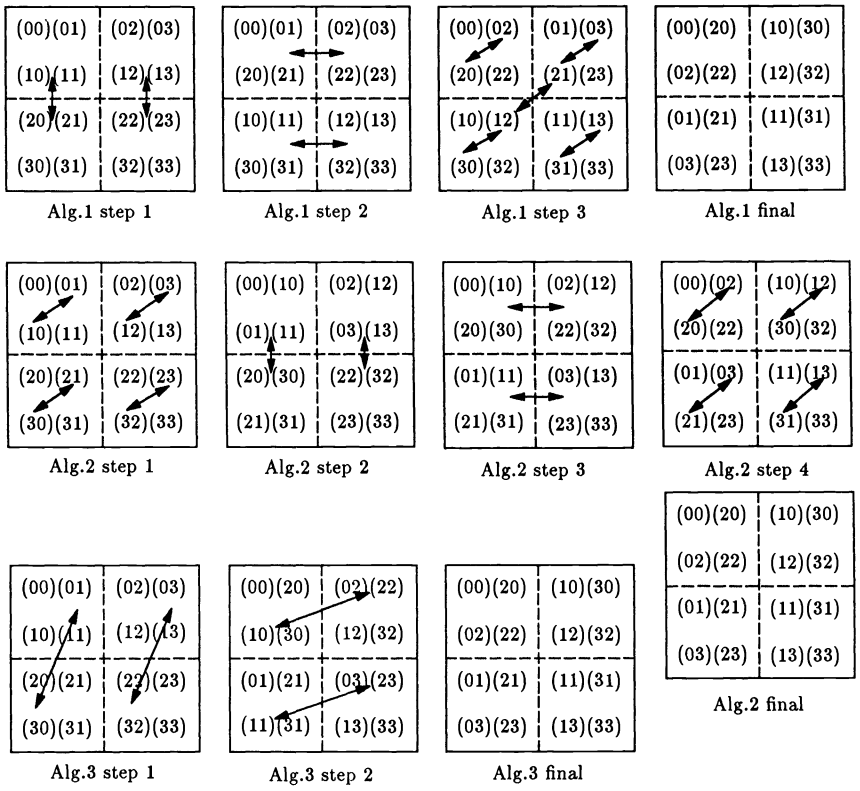


FIG. 5. *Three different algorithms to transpose a matrix from two-dimensional consecutive partitioning to two-dimensional cyclic partitioning.*

operations. Pipelining can be applied. For simplicity, we describe the nonpipelined version. As for the SPT algorithm, the combined algorithm is composed of $\frac{n}{2}$ iterations. Each iteration contains two routing steps. In iteration $i \in \{0, 1, \cdots, \frac{n}{2} - 1\}$, bits $\frac{n}{2} - i - 1$ of the row and column indices are changed by sending data through the corresponding dimensions. With the rows encoded in binary code and the columns in Gray code, matrix block (u, v) is stored in processor $(u||G(v))$ and matrix block (v, u) is stored in processor $(v||G(u))$. The direct transpose permutation is defined by exchanging data between processor $(u||G(v))$ and processor $(G^{-1}(G(v))||G(u))$, where $G^{-1}(\ )$ is the inverse Gray code.

During the first iteration, the upper right block $(0x_{n-2}x_{n-3}\cdots x_{\frac{n}{2}}||1x_{\frac{n}{2}-2}x_{\frac{n}{2}-3}\cdots x_0)$ and the lower left block $(1x_{n-2}x_{n-3}\cdots x_{\frac{n}{2}}||0x_{\frac{n}{2}-2}x_{\frac{n}{2}-3}\cdots x_0)$ are exchanged in two steps. Neither row nor column conversions for the two encodings affect iteration 0, because the Gray and binary codes have identical most significant bits. During the second iteration, the Gray code encoding of the column indices forces a horizontal exchange within the blocks for the second half of the block rows. The binary code encoding of the row indices forces a vertical exchange for the second half of the block columns. The transpose operation requires an antidiagonal exchange within all four blocks. The combined permutation pattern is shown in Fig. 6.

In general, the Gray code encoding of the columns causes a horizontal exchange within all the odd block rows with block rows numbered from 0. The binary code encoding causes a vertical exchange within all ith block columns such that the parity

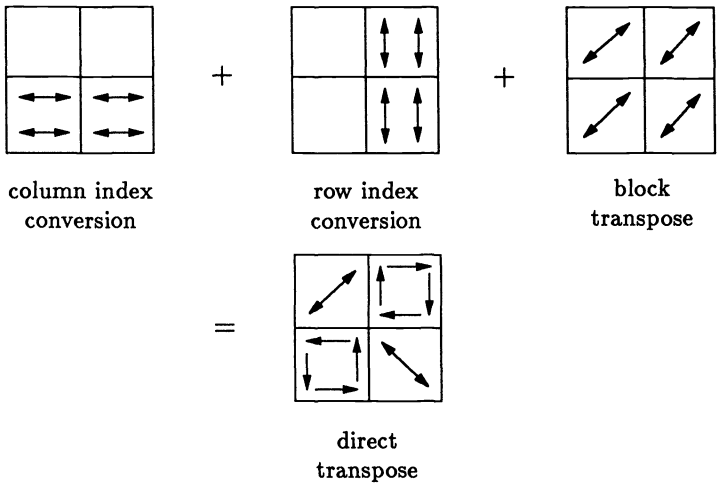


FIG. 6. *Transpose of a matrix stored by binary code encoding of row index and Gray code encoding of column index.*

of the binary encoding of i is odd. This can be proved from the conversion from binary code to Gray code proceeding from the most significant bit to the least significant bit (instead of a "low-order to high-order bit" conversion sequence [9]). Figure 7 shows the four iterations with $n = 8$, in which c means *clockwise rotation* and cc means *counterclockwise rotation.* The algorithm is presented below.

```
/* The second argument of "send" and "recv" represents the cube dimension */
/* and "buf" contains the data to be transposed initially. */
even-block-row := true;
even-parity-block-column := true;
for j := n/2 - 1 downto 0 do
        case (even-block-row, even-parity-block-column, bit j + n/2, bit j) of
                (TT00), (TT11), (FF01), (FF10):
                        recv (tmp, j + n/2); send (tmp, j);
                (TT01), (TT10), (FF00), (FF11), (TF01), (TF10), (FT00), (FT11):
                        send (buf, j + n/2); recv (buf, j);
                (TF00), (TF11), (FT01), (FT10):
                        send (buf, j); recv (buf, j + n/2);
        endcase
        even-block-row := (bit j + n/2 = 0);
        if (bit j = 1) then
                even-parity-block-column := not even-parity-block-column;
        endif
endfor
```

The above algorithm was implemented on the Intel iPSC. The results are shown in Fig. 13 in § 8.2.1, which discusses experiments.

To transpose a matrix stored by binary encoding of row and column indices into a transposed matrix with row and columns encoded in Gray code, a combined conversion-transpose algorithm similar to the one above can be applied to accomplish the task in n routing steps. The algorithm above needs only to be modified such

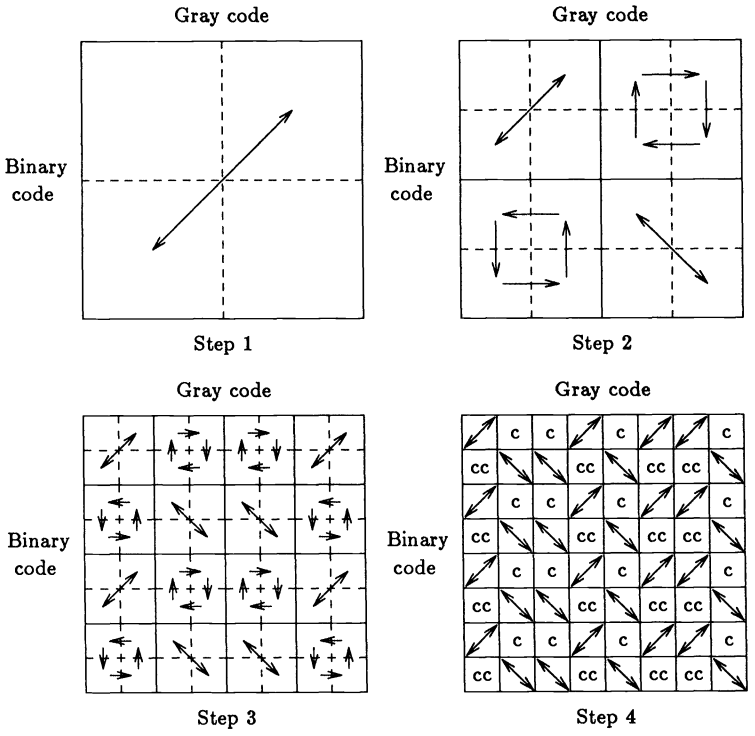


FIG. 7. *Transpose of a matrix stored by mixed encoding of rows and columns in an 8-cube.*

that the column operations are controlled by even-block-columns (instead of even-parity-block-columns). Similarly, to transpose a matrix with both row and columns encoded in Gray code into a transposed matrix with rows and columns encoded in binary code, the control of the row operations is changed from even-block-rows to even-parity-block-rows.

7. Using matrix transposition for other permutations. For $I = \phi$, and $|\mathcal{R}_b| = |\mathcal{R}_a| = n$, matrix transposition is an *all-to-all personalized communication.* An arbitrary permutation on an n-cube can be realized by all-to-all personalized communication twice, if the size of messages to be permuted is the same for all processors and at least N (per processor) [21], [20]. Since transposing a matrix with two-dimensional partitioning and $n_c = n_r$ is a permutation, we can also realize it by performing all-to-all personalized communication twice. However, the communication complexity is higher than that of the best transpose algorithm for the two-dimensional partitioning either for *one-port* communication, or for *n-port* communication.

The correspondence between cube dimensions for the *standard* exchange algorithm applied to matrix transposition is $f(i) = i$, $g(i) = i+\frac{n}{2}$, for all $i \in \{0, 1, \cdots, \frac{n}{2}-1\}$. By changing the exchange dimensions such that $f(i) = i$, $g(i) = n-1-i$, for all $i \in \{0, 1, \cdots, \frac{n}{2}-1\}$, a *bit-reversal* permutation is realized by the *general* exchange algorithm. A bit-reversal permutation is defined by

$$(x_{n-1}x_{n-2}\cdots x_0) \leftarrow (x_0x_1\cdots x_{n-1}).$$

DEFINITION 17. Define *dimension permutation* to be a permutation such that processor $(x_{n-1}x_{n-2}\cdots x_0)$ sends its data to processor $(x_{\delta(n-1)}x_{\delta(n-2)}\cdots x_{\delta(0)})$,

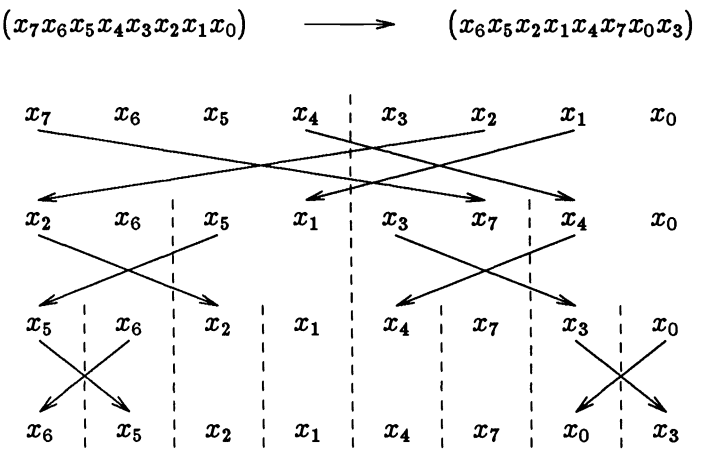


FIG. 8. *Realizing dimension permutation by performing* $\log n$ *steps of parallel swapping.*

where δ is a $\{0, 1, \cdots, n-1\}$ to $\{0, 1, \cdots, n-1\}$ permutation function.

DEFINITION 18. Define *parallel swapping* to be a dimension permutation such that the permutation function δ satisfies $\delta(\delta(i)) = i$, i.e., either $\delta(i) = i$ or $\delta(i) = j$, $\delta(j) = i$, $i \neq j$, for all $i \in \{0, 1, \cdots, n-1\}$

LEMMA 7.1. *Any dimension permutation can be realized by performing parallel swapping* $\lceil \log_2 n \rceil$ *times (note that* n *is the number of dimensions).*

Proof. First assume n is a power of 2. Arbitrarily partition the set of dimensions into two same-sized subsets, called S_1 and S_2, respectively. Let k be the cardinality of the set $\{i | i \in S_1, \delta(i) \in S_2\}$. Clearly, the cardinality of the set $\{i | i \in S_2, \delta(i) \in S_1\}$ is also k. Exchanging the k dimensions in S_1 with the corresponding k dimensions in S_2 can be done in one *parallel swapping* step. After this parallel swapping, there are two same-sized subsets which only require internal permutation. This permutation can be performed concurrently for the two subsets. Therefore, $\log n$ steps of *parallel swapping* suffice to realize the dimension permutation. For arbitrary n, we can add virtual elements such that the number of dimensions in the address field becomes a power of 2. □

Figure 8 shows an example of permuting eight dimensions by three steps of *parallel swappings.* Notice that k shuffle/unshuffle operations (left/right rotation k steps) fall in the *dimension permutation* class. There are $n!$ possible dimension permutations among $N!$ arbitrary permutations.

8. Experiments and implementation issues.

8.1. One-dimensional partitioning. The Intel iPSC effectively allows communication on only one port at a time. Hence, we choose to implement the one-dimensional transpose using the exchange algorithm. In our implementation we do not perform local shuffle operations in order to arrange the data to be exchanged into one block for the sake of reducing the number of start-ups, since the copying time on the Intel iPSC is significant. Copying 1024 single precision floating-point numbers (4 *k*bytes) takes about 37 milliseconds according to our measurements. The local array is partitioned into 2^j same-sized blocks during step j of the exchange algorithm. The odd or even blocks can either be sent directly to minimize the copy time, or copied into a buffer to reduce the number of start-ups. Figure 9 presents the measurements for unbuffered and buffered communication for rearrangement of consecutive to cyclic

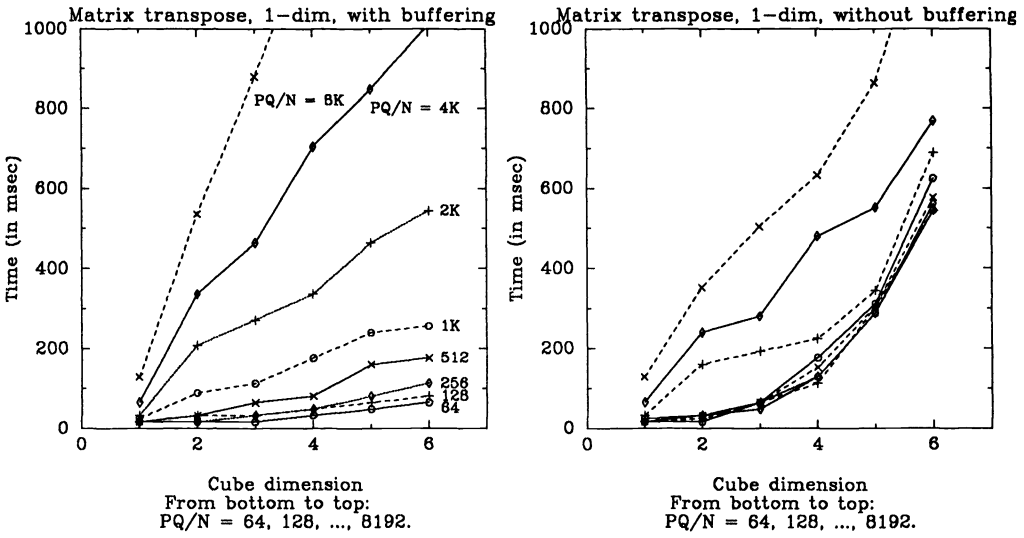


FIG. 9. *Measured times on the* Intel iPSC *for the transpose of a matrix, one-dimensional partitioning (or for conversion of consecutive to cyclic one-dimensional partitioning), encoded in binary code.*

partitioning.

The complexity of the unbuffered communication is easily found to be

$$T = n\frac{PQ}{2N}t_c + \left(N + \left\lceil \frac{PQ}{2B_mN} \right\rceil \min\left(n, \log_2 \left\lceil \frac{PQ}{B_mN} \right\rceil\right) - \frac{PQ}{B_mN}\right)\tau.$$

With buffered communication, messages may initially be larger than the buffer size, in which case they are sent directly. Small messages are buffered and the time for communication is

$$T = \left(\min\left(n, \log \left\lceil \frac{PQ}{B_mN} \right\rceil\right)\right)\left\lceil \frac{PQ}{2B_mN} \right\rceil + \min\left(N, \frac{PQ}{B_{\text{copy}}N}\right) - \min\left(N, \frac{PQ}{B_mN}\right) + \left\lceil \frac{PQ}{2B_mN} \right\rceil \max\left(0, n - \log \left\lceil \frac{PQ}{B_{\text{copy}}N} \right\rceil\right)\Bigg)\tau + n\frac{PQ}{2N}t_c + \frac{PQ}{N}\max\left(0, n - \log \left\lceil \frac{PQ}{B_{\text{copy}}N} \right\rceil\right)t_{\text{copy}},$$

where B_{copy} is the array size beyond which it is preferable with respect to performance to send without copying into a buffer. The complexity of the unbuffered communication grows linearly in the number of processors, i.e., exponentially in the number of cube dimensions, as shown in Fig. 9. The buffered communication grows linearly in the number of cube dimensions. For a low growth rate it is important to have a large buffer, to reduce the number of start-ups, and fast copy. With the times for copy of floating-point numbers and communication start-ups on the Intel iPSC the copy of 64 single-precision floating-point numbers (256 bytes) takes approximately the same time as one communication start-up. Hence, it is beneficial with respect to performance to send blocks of length at least 64 floating-point numbers without buffering. Figure 10 shows the improvement in performance with optimum buffering compared to the unbuffered communication. Note that for sufficiently small cubes (or large data sets) the time required by the two schemes coincide.

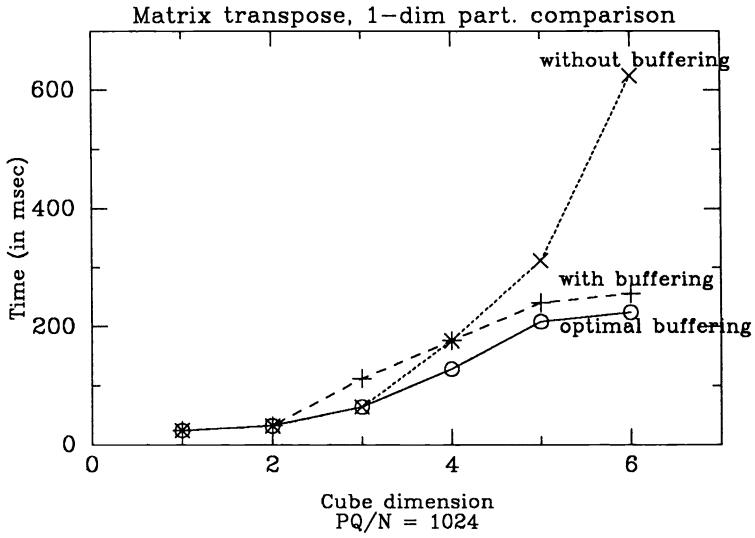


FIG. 10. *The effect of optimum buffering on performance for matrix transpose on the* Intel iPSC.

On the iPSC, it is also possible to realize the *all-to-all personalized communication* by calling the iPSC router $2(N-1)$ times. However, the measured times of this are always inferior to that of the optimum buffering algorithm. The difference is from a factor of 5 to two orders of magnitude depending on the matrix size and cube size as observed in [13].

8.2. Two-dimensional partitioning.

8.2.1. The Intel iPSC. We have implemented algorithm SPT as a step by step procedure. Pipelining is not possible. Moreover, on the Intel iPSC it is necessary to rearrange two-dimensional arrays into one-dimensional arrays before sending. Since the copy time is significant we arrive at an estimate for the time of a two-dimensional transpose of $T = (\frac{PQ}{N}t_c + \lceil \frac{PQ}{B_m N} \rceil \tau)n + 2\frac{PQ}{N}t_{\text{copy}}$. The growth rate is proportional to the number of matrix elements. There is an exponential decay as well as a linear increase in the number of cube dimensions. Figure 11 shows measured values for the copy time, the communication time and the total time for a 2-cube and a 6-cube. As expected, the copy time for the 6-cube is lower than that for the 2-cube. Also, the communication time is essentially determined by the number of start-ups, which for the 6-cube remains the same for $PQ \leq 64$ *k*bytes.

Figure 12(a) shows the total transpose time as a function of the number of cube dimensions and matrix size. For small matrices the number of communication start-ups dominates and the total time increases with the number of cube dimensions, but as the matrix size increases the transpose time decreases with increased cube size.

On the Intel iPSC it is also possible to carry out the transpose operation by a direct send to the final destination. Figure 12(b) gives the times measured for matrix transpose using the routing logic alone. As the cube size increases the two-dimensional transpose algorithm yields a significantly better performance than the transpose time offered by the routing logic.

The time for matrix transposition with simultaneous conversion from Gray code to binary code conversion is shown in Fig. 13. It is assumed that rows and columns

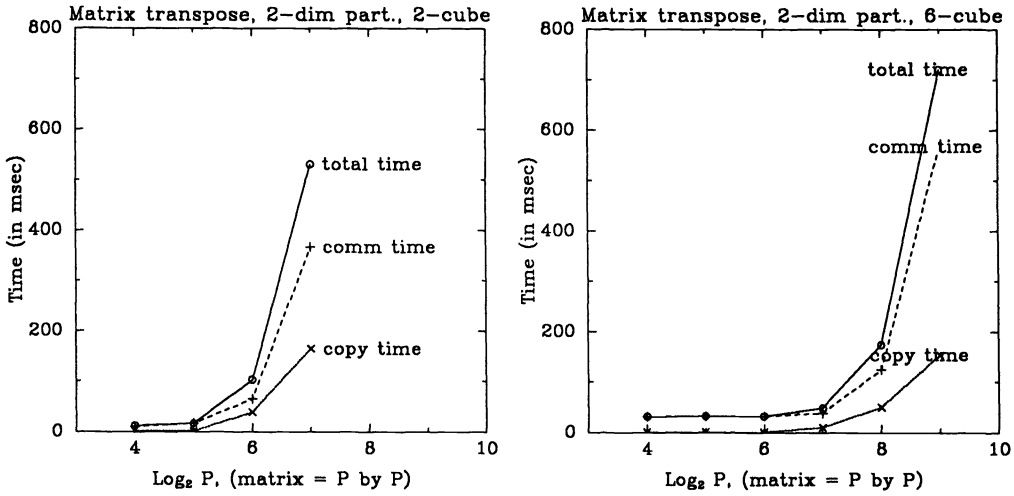


FIG. 11. *Performance measurements for a two-dimensional matrix transpose on the* Intel iPSC.

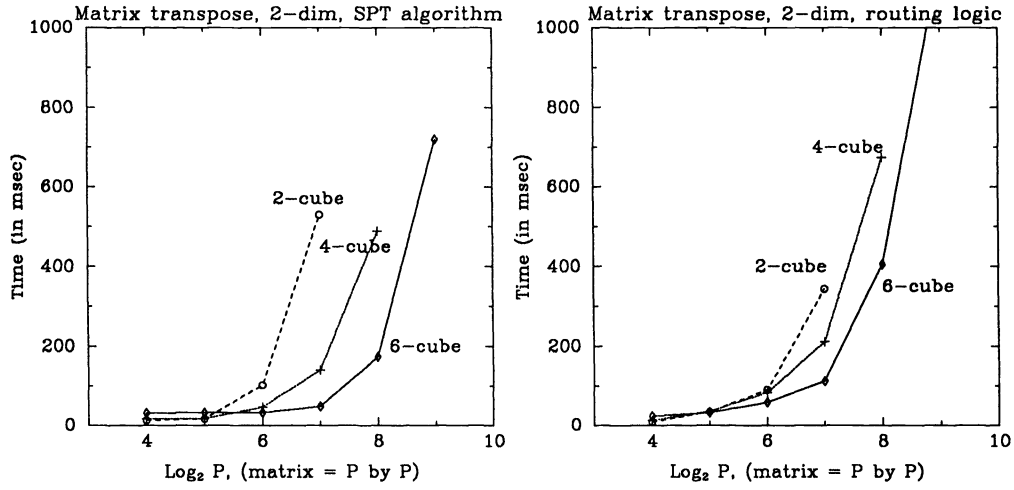


FIG. 12. *Measured times for a two-dimensional matrix transpose on the* Intel iPSC *using the* SPT *algorithm without pipelining* (a) *and using routing logic* (b).

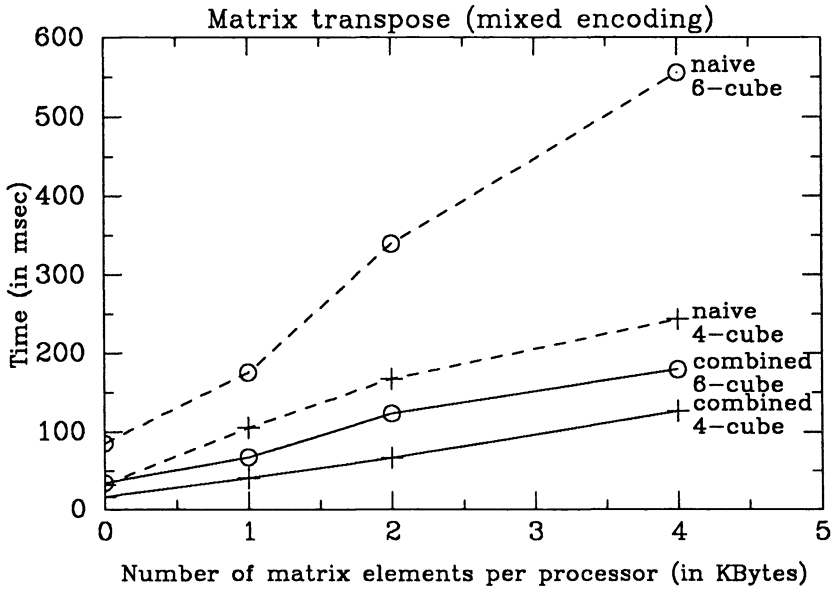


FIG. 13. *Measured times of transposing a matrix stored by mixed encoding of rows and columns by the naive and combined algorithms on the* Intel iPSC.

have different encoding schemes. The figure compares the $2n-2$ steps naive algorithm and the n steps combined algorithm.

8.2.2. The Connection Machine. We have also implemented the matrix transpose operation on the Connection Machine. It has a bit-serial, pipelined communication system. The recursive algorithm does not exploit this feature, but the routing logic does. Figure 14 shows the transpose time using the routing logic. Each processor holds one matrix element (32 bits). Figure 15 shows the transpose times for various number of matrix elements per processor, and for various number of processors. Figure 16 shows the transpose times for two fixed sized matrices on various sizes of the Connection Machine.

9. Comparison and conclusion. It is of interest to compare the times for matrix transpose based on a one-dimensional partitioning and a two-dimensional partitioning. We now compare the complexity estimate for the two-dimensional transpose

$$T^{2d} = \left(\frac{PQ}{N}t_c + \left\lceil \frac{PQ}{B_m N} \right\rceil \tau\right) n + 2\frac{PQ}{N}t_{\text{copy}}$$

with that for the one-dimensional transpose

$$T^{1d} = \left(\min\left(n, \log\left\lceil \frac{PQ}{B_m N}\right\rceil\right)\left\lceil \frac{PQ}{2B_m N}\right\rceil + \min\left(N, \frac{PQ}{B_{\text{copy}}N}\right) - \min\left(N, \frac{PQ}{B_m N}\right) + \left\lceil \frac{PQ}{2B_m N}\right\rceil \max\left(0, n - \log\left\lceil \frac{PQ}{B_{\text{copy}}N}\right\rceil\right)\right)\tau + n\frac{PQ}{2N}t_c + \frac{PQ}{N}\max\left(0, n - \log\left\lceil \frac{PQ}{B_{\text{copy}}N}\right\rceil\right)t_{\text{copy}}.$$

We have assumed that *one exchange* takes the same time as one send *or* one receive for *one-port* communication throughout the paper. With this model, the

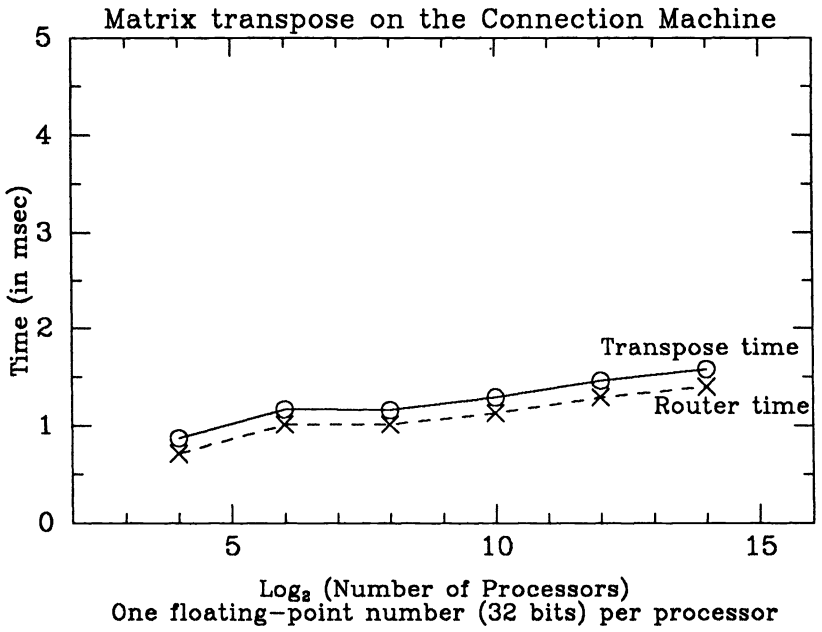


FIG. 14. *Matrix transpose on the Connection Machine. One element per processor.*

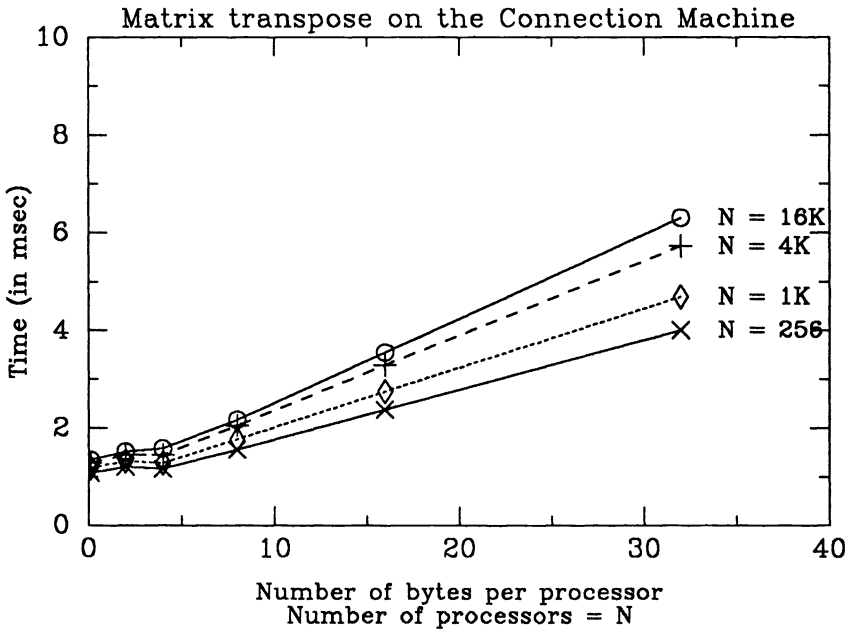


FIG. 15. *Matrix transpose on the Connection Machine. Multiple elements per processor.*

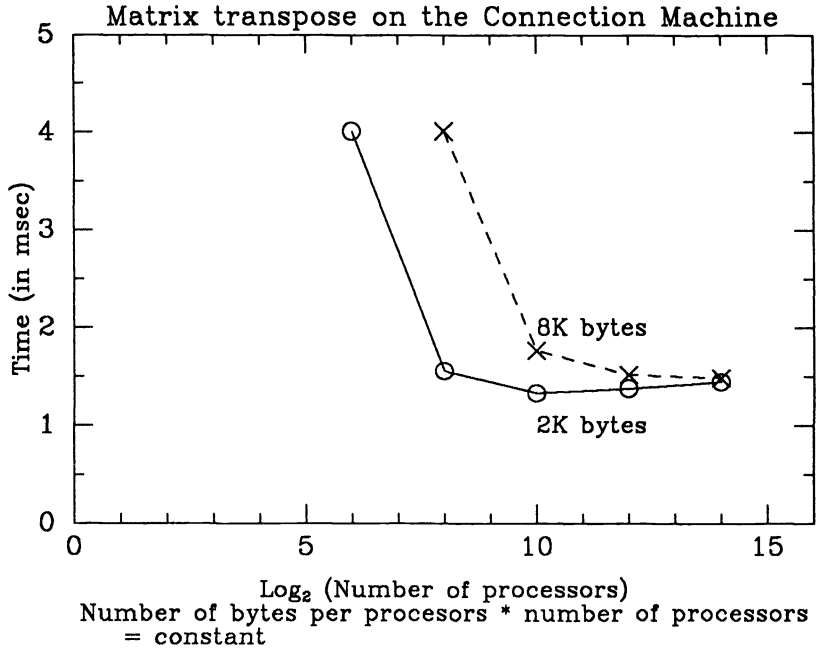


FIG. 16. *Matrix transpose on the Connection Machine as a function of the machine size.*

time for data transfers for the one-dimensional transpose is half of that of the two-dimensional transpose. If copy time is negligible, i.e., the time to copy B_m data is much less than a start-up time, then the number of start-ups for one-dimensional transpose is a factor of $\frac{1}{2}$ to 1 of that for the two-dimensional transpose. The factor $\frac{1}{2}$ applies for $\frac{PQ}{2N} \geq B_m$. By considering the copy time, we have two extreme cases. If $\frac{PQ}{N^2} \geq B_m$, the number of start-ups for the one-dimensional transpose is half of that for the two-dimensional transpose. If $\frac{PQ}{N} \leq B_{\text{copy}}$, it can be shown that the number of start-ups for the one-dimensional transpose is at most twice that for the two-dimensional transpose. In general, it can be shown that the number of start-ups for the one-dimensional transpose is a factor of $\frac{1}{2}$ to $\frac{B_m}{2B_{\text{copy}}} + \frac{1}{2}$ (which is 2.5 for the Intel iPSC) of that for the two-dimensional transpose.

If the communication is restricted to one send *or* one receive at a time, the time for data transfers and the number of start-ups increase by a factor of two for the one-dimensional transpose. However, the complexity for the two-dimensional transpose remains the same. Therefore, the complexity of the two-dimensional transpose will be lower, or the same, as that of the one-dimensional transpose by a factor of 1 to $\frac{B_m}{B_{\text{copy}}} + 1$. Figure 17 gives the experimental result on the Intel iPSC.

With concurrent communication on multiple ports the transfer time for the two-dimensional partitioning decreases exponentially in the number of cube dimensions, but for the optimum packet size the number of start-ups is higher than for the one-dimensional partitioning. From the complexity estimates (one-dimensional partitioning)

$$T_{\min}^{1d} = \frac{PQ}{2N} t_c + n\tau$$

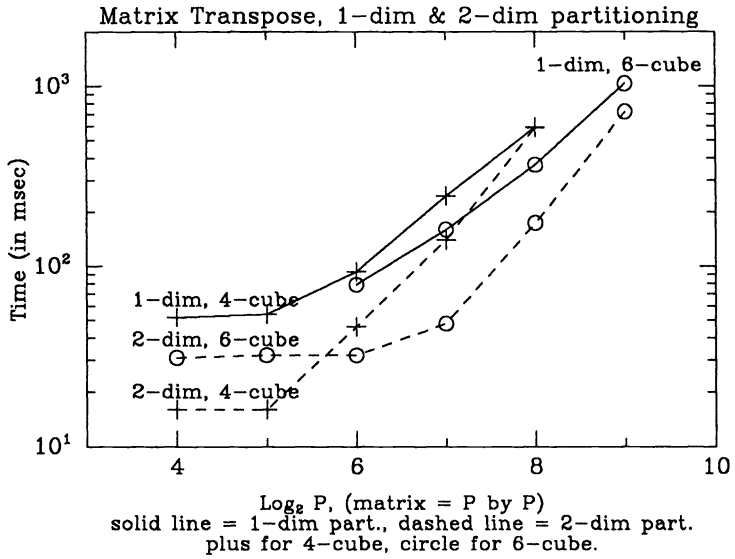


FIG. 17. *Comparison of the matrix transpose operation of one- and two-dimensional partitioned matrices on the* Intel iPSC.

and

$$T_{\min}^{2d} = \begin{cases} (n+1)\tau + \frac{n+1}{2n}\frac{PQ}{N}t_c & \text{if } n \geq \sqrt{\frac{PQt_c}{N\tau}} \text{ approximately,} \\ (\frac{n}{2}+3)\tau + \frac{n+6}{2n+8}\frac{PQ}{N}t_c & \text{if } \sqrt{\frac{PQt_c}{2N\tau}} < n \leq \sqrt{\frac{PQt_c}{N\tau}} \text{ approximately and } \frac{n}{2} \text{ is even,} \\ (\frac{n}{2}+2)\tau + \frac{n+4}{2n+4}\frac{PQ}{N}t_c & \text{if } \sqrt{\frac{PQt_c}{2N\tau}} < n \leq \sqrt{\frac{PQt_c}{N\tau}} \text{ approximately and } \frac{n}{2} \text{ is odd,} \\ (\sqrt{\tau} + \sqrt{\frac{PQt_c}{2N}})^2 & \text{if } n \leq \sqrt{\frac{PQt_c}{2N\tau}}. \end{cases}$$

The optimum packet size is

$$B_{\text{opt}} = \begin{cases} \lceil \frac{PQ}{N(n+4)} \rceil & \text{for even } \frac{n}{2} \text{ and } n > \sqrt{\frac{PQt_c}{2N\tau}}, \\ \lceil \frac{PQ}{N(n+2)} \rceil & \text{for odd } \frac{n}{2} \text{ and } n > \sqrt{\frac{PQt_c}{2N\tau}}, \\ \sqrt{\frac{PQ\tau}{2Nt_c}} & \text{for } n \leq \sqrt{\frac{PQt_c}{2N\tau}}. \end{cases}$$

For $n \geq \sqrt{\frac{PQt_c}{N\tau}}$, the one-dimensional partitioning always yields a lower complexity than the two-dimensional partitioning. The difference is about one start-up time unless the cube is very small. For $\sqrt{\frac{PQt_c}{2N\tau}} < n \leq \sqrt{\frac{PQt_c}{N\tau}}$, the break even point (ignoring copy) can be computed to be

$$N \approx c\frac{r}{\log^2 r},$$

where $\frac{1}{2} < c < 1$ and $r = \frac{PQt_c}{\tau}$. For $n \leq \sqrt{\frac{PQt_c}{2N\tau}}$, the one-dimensional partitioning always yields a lower complexity than the two-dimensional partitioning.

In summary, if the copy time is ignored and communication is restricted to one port at a time, then the one-dimensional partitioning always yields a lower complexity

than the two-dimensional partitioning. If the copy time is included then the two-dimensional partitioning yields a lower complexity for a sufficiently large cube. With concurrent communication on all ports the *Spanning Balanced n-Tree* (SBnT) routing can be used for the one-dimensional partitioning, and the copy times for one- and two-dimensional partitioning should be comparable. The one-dimensional partitioning yields a lower complexity for a cube dimension n satisfying $n \geq \sqrt{\frac{PQt_c}{N\tau}}$ or $n \leq \sqrt{\frac{PQt_c}{2N\tau}}$.

In comparing the Intel iPSC with the Connection Machine we conclude that the latter performs a transpose about two orders of magnitude faster.

Acknowledgments. The authors express their appreciation for the comments, made by one of the referees, which helped to improve the presentation of the results over the original version of this paper.

REFERENCES

[1] J. O. EKLUNDH, *A fast computer method for matrix transposing*, IEEE Trans. Comput., 21 (1972), pp. 801–803.

[2] G. C. FOX AND W. FURMANSKI, *Optimal communication algorithms on hypercube*, Tech. Report, California Institute of Technology, July 1986.

[3] W. D. HILLIS, *The Connection Machine*, MIT Press, Cambridge, MA, 1985.

[4] C. HO AND S. L. JOHNSSON, *Dimension permutation on Boolean cubes*, Tech. Report, Dept. of Computer Science, Yale Univ., New Haven, CT, to appear.

[5] ———, *Distributed routing algorithms for broadcasting and personalized communication in hypercubes*, in 1986 International Conf. on Parallel Processing, IEEE Computer Society, 1986, pp. 640–648. Tech. Report YALEU/DCS/RR-483, May 1986.

[6] ———, *Spanning balanced trees in Boolean cubes*, Tech. Report YALEU/DCS/RR-508, Dept. of Computer Science, Yale Univ., New Haven, CT, January 1987.

[7] S. L. JOHNSSON, *Band matrix systems solvers on ensemble architectures*, in Algorithms, Architecture, and the Future of Scientific Computation, Univ. of Texas Press, Austin, TX, 1985. (Tech. Report YALEU/DCS/RR-388, Yale Univ., New Haven, CT, May 1985).

[8] ———, *Communication efficient basic linear algebra computations on hypercube architectures*, J. Parallel Distributed Comput., 4 (1987), pp. 133–172. (Tech. Report YALEU/DCS/RR-361, Yale Univ., New Haven, CT, January 1985).

[9] ———, *Data permutations and basic linear algebra computations on ensemble architectures*, Tech. Report YALEU/DCS/RR-367, Dept. of Computer Science, Yale Univ., New Haven, CT, February 1985.

[10] ———, *Odd-even cyclic reduction on ensemble architectures and the solution tridiagonal systems of equations*, Tech. Report YALE/DCS/RR-339, Dept. of Computer Science, Yale Univ., October 1984.

[11] ———, *Solving narrow banded systems on ensemble architectures*, ACM Trans. Math. Software, 11 (1985), pp. 271–288. (Tech. Report YALEU/DCS/RR-418, Yale Univ., New Haven, CT, November 1984).

[12] ———, *Solving tridiagonal systems on ensemble architectures*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 354–392. (Tech. Report YALEU/DCS/RR-436, Yale Univ., New Haven, CT, November 1985).

[13] S. L. JOHNSSON AND C. HO, *Multiple tridiagonal systems, the alternating direction method, and Boolean cube configured multiprocessors*, Tech. Report YALEU/DCS/RR-532, Dept. of Computer Science, Yale Univ., New Haven, CT, June 1987.

[14] ———, *Spanning graphs for optimum broadcasting and personalized communication in hypercubes*, Tech. Report YALEU/DCS/RR-500, Dept. of Computer Science, Yale Univ., New Haven, CT, November 1986. To appear in IEEE Trans. Computers.

[15] O. A. MCBRYAN AND E. F. V. DE VELDE, *Hypercube algorithms and implementations*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s227–s287.

[16] E. M. REINGOLD, J. NIEVERGELT, AND N. DEO, *Combinatorial Algorithms*, Prentice-Hall, Englewood Cliffs. NJ, 1977.

[17] Y. SAAD AND M. H. SCHULTZ, *Data communication in hypercubes*, Tech. Report YALEU/DCS/RR-428, Dept. of Computer Science, Yale Univ., New Haven, CT, October 1985.

[18] ———, *Topological properties of hypercubes*, Tech. Report YALEU/DCS/RR-389, Dept. of Computer Science, Yale Univ., New Haven, CT, June 1985.

[19] H. S. STONE, *Parallel processing with the perfect shuffle*, IEEE Trans. Comput., 20 (1971), pp. 153–161.

[20] Q. F. STOUT AND B. WAGER, *Intensive hypercube communication I: prearranged communication in link-bound machines*, Tech. Report CRL-TR-9-87, Computing Research Lab., Univ. of Michigan, Ann Arbor, MI, 1987.

[21] ———, *Passing messages in link-bound hypercubes*, in Hypercube Multiprocessors 1987, M. T. Heath, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 4, October 1988

CLASSIFICATIONS OF NONNEGATIVE MATRICES USING DIAGONAL EQUIVALENCE*

DANIEL HERSHKOWITZ†, URIEL G. ROTHBLUM‡, AND HANS SCHNEIDER§

Abstract. This article studies matrices A that are positively diagonally equivalent to matrices that, for given positive vectors u, v, r, and c, map u into r, and where A^T map v into c. The problem is reduced to scaling a matrix for given row sums and column sums, and applying known results for the latter. Further classifications that use these results are investigated.

Key words. diagonal equivalence, nonnegative matrices, classification of nonnegative matrices

AMS(MOS) subject classifications. 65F35, 15A21, 15A48

1. Introduction. The problem of examining matrices that map a given n-dimensional vector into a given m-dimensional vector underlines many important issues in linear algebra. For example, the assertion that the row sums and/or the column sums of a matrix A are given by vectors r and c, respectively, means that A maps e into r and/or that A^T maps e into c, where e denotes the vector of appropriate dimension all of whose coordinates are 1. Also, the statement that a square matrix A has a right eigenvector u and a left eigenvector v corresponding to a nonzero eigenvalue λ, means that A/λ maps u into u and that A^T/λ maps v into v. Another example is the statement that the null space of a matrix A contains the vector x, which means that A maps x into the zero vector.

The purpose of this paper is to study matrices that are positively diagonally equivalent to nonnegative matrices A that map u into r, and where A^T map v into c for given positive vectors u, v, r, and c. We show that, in general, the set of such matrices can be represented as the set of matrices that are positively diagonally equivalent to nonnegative matrices having prespecified row sums and column sums. We then use a known characterization of the latter class to obtain a characterization of the former class. We also characterize matrices in the intersection, as well as in the union of these classes, over all possible choices of the vectors u, v, r, and c for which these sets are nonempty. We also obtain a special characterization for the eigenvector problem, where $m = n$, $u = r$, and $v = c$.

2. Notation and definitions.

Notation 2.1. Let m and n be positive integers. We denote by

$\langle n \rangle$, the set $\{1, 2, \cdots, n\}$;

R_{+0}^{mn}, the set of all nonnegative $m \times n$ matrices;

R_{+}^{n}, the set of all positive $n \times 1$ column vectors;

e_n, the $n \times 1$ column vector all of whose components are 1.

Notation 2.2. For a set α we denote by $|\alpha|$ the cardinality of α.

Notation 2.3. Let A be an $m \times n$ matrix and let α and β be nonempty subsets of $\langle m \rangle$ and $\langle n \rangle$, respectively. We denote by $A[\alpha|\beta]$ the submatrix of A whose rows and

* Received by the editors August 10, 1987; accepted for publication (in revised form) January 13, 1988. The research of the first and second authors was supported by grant 85-00153 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel. The second author's research was also supported by National Science Foundation grants DMS-8521521 and ECSE-18971. The third author's research was supported by National Science Foundation grant ECS-83-10213.

† Mathematics Department, Technion–Israel Institute of Technology, Haifa 32000, Israel.

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

§ Mathematics Department, University of Wisconsin, Madison, Wisconsin 53706.

columns are indexed by the elements of α and β, respectively, in their natural order. Also, we denote by α' and β' the sets $\langle n\rangle\backslash\alpha$ and $\langle n\rangle\backslash\beta$, respectively.

Notation 2.4. Let x be an $n \times 1$ column vector and let $\alpha \subseteq \langle n\rangle$. We denote by x_α the subvector of x whose coordinates are indexed by the elements of α.

Notation 2.5. Let m and n be positive integers, let $u, c \in R_+^{\mathbf{n}}$, and let $v, r \in R_+^{\mathbf{m}}$. We denote

$$F_{mn}(u,v,r,c) = \{A \in R_{+0}^{\mathbf{mn}} : Au = r, v^{\mathbf{T}}A = c^{\mathbf{T}}\},$$

$$S_{mn}(r,c) = F_{mn}(e_n, e_m, r, c).$$

In the case that $m = n$ we denote

$$E_{nn}(u,v) = F_{nn}(u,v,u,v).$$

Remark 2.6. Observe that $S_{mn}(r, c)$ is the set of all $m \times n$ nonnegative matrices with row sums $r_1, \cdots, r_m$ and column sums $c_1, \cdots, c_n$. The set $E_{nn}(u, v)$ consists of all $n \times n$ nonnegative matrices with eigenvalue 1, where u and v are the corresponding right and left eigenvectors.

Notation 2.7. Let u be a vector. We denote by U the diagonal matrix whose diagonal elements are $u_1, \cdots, u_n$. Similar relations hold between v, r, c and V, R, C respectively.

DEFINITION 2.8. A diagonal matrix is said to be *positive diagonal* if it has positive diagonal elements.

DEFINITION 2.9. Let A and B be $m \times n$ matrices. We say that A and B are *positively diagonally equivalent* if there exists positive diagonal matrices $D \in R_{+0}^{\mathbf{mm}}$ and $E \in R_{+0}^{\mathbf{nn}}$ such that $A = DBE$.

Notation 2.10. Let $u, c \in R_+^{\mathbf{n}}$ and let $v, r \in R_+^{\mathbf{m}}$. We denote the set of all $B \in R_{+0}^{\mathbf{mn}}$ such that B is positively diagonally equivalent to some $A \in F_{mn}(u, v, r, c)$ by $F_{mn}^*(u, v, r, c)$. Also, we use the following notation:

$$S_{mn}^*(r,c) \equiv F_{mn}^*(e_n, e_m, r, c).$$

and in the case that $m = n$

$$E_{nn}^*(u,v) \equiv F_{nn}^*(u,v,u,v).$$

Notation 2.11. Let A and B be $m \times n$ matrices. We denote by $A \circ B$ the Hadamard product of A and B, viz., the $m \times n$ matrix C such that $c_{ij} = a_{ij}b_{ij}$, $i \in \langle m\rangle, j \in \langle n\rangle$. In particular, this notation applies when A and B are vectors. Obviously, the Hadamard product is commutative.

DEFINITION 2.12. An $m \times n$ matrix A is said to be *chainable* if it has no zero row or column, and if for every pair of nonempty proper subsets α and β of $\langle m\rangle$ and $\langle n\rangle$, respectively, $A[\alpha|\beta] = 0$ implies $A[\alpha'|\beta'] \neq 0$.

DEFINITION 2.13. Let m and n be positive integers, let $\alpha_1, \cdots, \alpha_p$ be nonempty pairwise disjoint subsets of $\langle m\rangle$ such that $\cup_{i=1}^p \alpha_i = \langle m\rangle$, and let $\beta_1, \cdots, \beta_p$ be nonempty pairwise disjoint subsets of $\langle n\rangle$ such that $\cup_{i=1}^p \beta_i = \langle n\rangle$. An $m \times n$ matrix A is said to be a (rectangular) *direct sum of* $A[\alpha_1|\beta_1], \cdots, A[\alpha_p|\beta_p]$ if $A[\alpha_i|\beta_j] = 0$ for all $i, j \in \langle p\rangle$, $i \neq j$.

We comment that every rectangular matrix having no zero row or zero column is a (rectangular) direct sum of chainable matrices $A[\alpha_i, \beta_i]$ for some sets $\alpha_1, \cdots, \alpha_p$ that partition $\langle m\rangle$, and for sets $\beta_1, \cdots, \beta_p$ that partition $\langle n\rangle$.

3. The classes $F_{mn}^*(u, v, r, c)$, $S_{mn}^*(r, c)$, and $E_{mn}^*(u, v)$.

LEMMA 3.1. *Let $A \in R_{+0}^{\mathbf{mn}}$, let $u, c \in R_+^{\mathbf{n}}$ and let $v, r \in R_+^{\mathbf{m}}$. Then $A \in F_{mn}(u, v, r, c)$ if and only if $VAU \in S_{mn}(r \circ v, c \circ u)$.*

Proof. The statement $A \in F_{mn}(u, v, r, c)$ means

$$AUe_n = Re_m, \qquad e_m^{\mathrm{T}} VA = e_n^{\mathrm{T}} C, \tag{3.2}$$

while the statement $VAU \in S_{mn}(r \circ v, c \circ u)$ means

$$VAUe_n = VRe_m, \qquad e_m^{\mathrm{T}} VAU = e_n^{\mathrm{T}} CU. \tag{3.3}$$

The equivalence of (3.2) and (3.3) is clear. □

COROLLARY 3.4. *Let $u, c \in R_+^{\mathbf{n}}$ and let $v, r \in R_+^{\mathbf{m}}$. If $F_{mn}(u, v, r, c)$ is nonempty then $v^{\mathrm{T}} r = c^{\mathrm{T}} u$.*

Proof. The result follows directly from Lemma 3.1 and the corresponding standard result concerning the transportation problem. □

COROLLARY 3.5. *Let $A \in R_{+0}^{\mathbf{mn}}$, let $u, c \in R_+^{\mathbf{n}}$, and let $v, r \in R_+^{\mathbf{m}}$. Then $A \in F_{mn}^*(u, v, r, c)$ if and only if $A \in S_{mn}^*(r \circ v, c \circ u)$.*

COROLLARY 3.6. *Let $A \in R_{+0}^{\mathbf{nn}}$ and let $u, v \in R_+^{\mathbf{n}}$. Then $A \in E_{nn}^*(u, v)$ if and only if $A \in S_{mn}^*(u \circ v, u \circ v)$.*

The following theorem is proved in [3] as Theorems 3.9 and 4.1. We state it here in a slightly different way.

THEOREM 3.7. *Let $A \in R_{+0}^{\mathbf{mn}}$ have no zero row or zero column, let $c \in R_+^{\mathbf{n}}$, and let $r \in R_+^{\mathbf{m}}$. Then we have the following:*

(i) *When A is chainable, then $A \in S_{mn}^*(r, c)$ if and only if for every pair of nonempty proper subsets α and β of $\langle m \rangle$ and $\langle n \rangle$ we have*

$$A[\alpha|\beta] = 0 \quad \text{and} \quad A[\alpha'|\beta'] \neq 0 \Rightarrow \sum_{i \in \alpha} r_i < \sum_{i \in \beta'} c_i.$$

In this case, there exist unique (up to scalar multiplication) positive diagonal matrices D and E such that $DAE \in S_{mn}(r, c)$.

(ii) *$A \in S_{mn}^*(r, c)$ if and only if A is a direct sum of chainable matrices $A[\alpha_i|\beta_i]$, $i = 1, \cdots, p$, such that*

$$A[\alpha_i|\beta_i] \in S_{|\alpha_i||\beta_i|}^*(r_{\alpha_i}, c_{\beta_i}), \qquad i \in \langle p \rangle.$$

(iii) *If $A \in S_{mn}^*(r, c)$ then there exists a unique matrix in $S_{mn}(r, c)$ which is positively diagonally equivalent to A.*

Remark 3.8. Statement (iii) in Theorem 3.7 follows immediately from statement (ii). Observe that in statement (iii) we do not assert uniqueness of the positive diagonal matrices D and E such that $DAE \in S_{mn}(r, c)$, but the uniqueness of the matrix DAE.

We now use our results in order to generalize Theorem 3.7. The following result also generalizes Theorem 3.10 of [1] and Theorem 3.2 of [2].

THEOREM 3.9. *Let $A \in R_{+0}^{\mathbf{mn}}$ have no zero row or zero column, let $u, c \in R_+^{\mathbf{n}}$, and let $v, r \in R_+^{\mathbf{m}}$. Then we have the following:*

(i) *When A is chainable then $A \in F_{mn}^*(u, v, r, c)$ if and only if for every pair of nonempty proper subsets α and β of $\langle m \rangle$ and $\langle n \rangle$ we have*

$$A[\alpha|\beta] = 0 \quad \text{and} \quad A[\alpha'|\beta'] \neq 0 \Rightarrow v_\alpha^{\mathrm{T}} r_\alpha < c_{\beta'}^{\mathrm{T}} u_{\beta'}.$$

In this case, there exist unique (up to scalar multiplication) positive diagonal matrices D and E such that $DAE \in F_{mn}(u, v, r, c)$.

(ii) *$A \in F_{mn}^*(u, v, r, c)$ if and only if A is a direct sum of chainable matrices $A[\alpha_i|\beta_i]$, $i = 1, \cdots, p$, such that*

$$A[\alpha_i|\beta_i] \in F_{|\alpha_i||\beta_i|}^*(u_{\beta_i}, v_{\alpha_i}, r_{\alpha_i}, c_{\beta_i}), \qquad i \in \langle p \rangle.$$

(iii) *If* $A \in F^*_{mn}(u, v, r, c)$ *then there exists a unique matrix in* $F_{mn}(u, v, r, c)$ *which is positively diagonally equivalent to* A.

Proof. The assertion follows directly from Corollary 3.5 and Theorem 3.7. □

In view of Corollary 3.4, statements (i) and (ii) of Theorem 3.9 can be combined and restated as Theorem 3.10.

THEOREM 3.10. *Let* $A \in R^{mn}_{+0}$, *have no zero row or zero column, let* $u, c \in R^n_+$, *and let* $v, r \in R^m_+$. *Then* $A \in F^*_{mn}(u, v, r, c)$ *if and only if for every pair of nonempty proper subsets* α *and* β *of* $\langle m \rangle$ *and* $\langle n \rangle$, *respectively, we have*

$$A[\alpha|\beta] = 0 \quad \text{and} \quad A[\alpha'|\beta'] \neq 0 \Rightarrow v_\alpha^T r_\alpha < c_{\beta'}^T u_{\beta'},$$

$$A[\alpha|\beta] = 0 \quad \text{and} \quad A[\alpha'|\beta'] = 0 \Rightarrow v_\alpha^T r_\alpha = c_{\beta'}^T u_{\beta'}.$$

4. The classes $\cap F^*_{mn}$, $\cup F^*_{mn}$, $\cap S^*_{mn}$, $\cup S^*_{mn}$, $\cap E^*_{nn}$, and $\cup E^*_{nn}$.

Notation 4.1. Let m and n be positive integers. We denote the following:

$$\cap F^*_{mn} = \bigcap_{\substack{u,c \in R^n_+ \\ v,r \in R^m_+ \\ u^T c = v^T r}} F^*_{mn}(u, v, r, c),$$

$$\cup F^*_{mn} = \bigcup_{\substack{u,c \in R^n_+ \\ v,r \in R^m_+}} F^*_{mn}(u, v, r, c),$$

$$\cap S^*_{mn} = \bigcap_{\substack{c \in R^n_+ \\ r \in R^m_+ \\ e_n^T c = e_m^T r}} S^*_{mn}(r, c),$$

$$\cup S^*_{mn} = \bigcup_{\substack{c \in R^n_+ \\ r \in R^m_+}} S^*_{mn}(r, c),$$

$$\cap E^*_{nn} = \bigcap_{u,v \in R^n_+} E^*_{nn}(u, v),$$

$$\cup E^*_{nn} = \bigcup_{u,v \in R^n_+} E^*_{nn}(u, v).$$

THEOREM 4.2. *Let* $A \in R^{mn}_{+0}$. *Then we have the following*:

(i) $A \in \cap F^*_{mn}$ *if and only if* A *has no zero entries.*

(ii) $A \in \cup F^*_{mn} \setminus \cap F^*_{mn}$ *if and only if* A *has at least one zero entry but there is no zero row or zero column in* A.

(iii) $A \notin \cup F^*_{mn}$ *if and only if* A *has at least one zero row or zero column.*

Proof. (i) Let $A \in R^{mn}_{+0}$. If A has no zero entries then Theorem 3.10 immediately implies that $A \in \cap F^*_{mn}$. Conversely, we show that if $a_{ij} = 0$ for some $i \in \langle m \rangle$ and $j \in \langle n \rangle$, then $A \notin \cap F^*_{mn}$. We choose $u, c \in R^n_+$ with $u_j c_j = \frac{2}{3}$ and $u^T c = 1$, and $v, r \in R^m_+$ with $v_i r_i = \frac{2}{3}$ and $v^T r = 1$. Then for $\alpha = \{i\}$ and $\beta = \{j\}$ we have that

$$v_\alpha^T r_\alpha = \tfrac{2}{3} > \tfrac{1}{3} = u^T c - u_j c_j = u_\beta^T c_{\beta'}.$$

Since $A[\alpha|\beta] = 0$ it now follows from Theorem 3.10 that $A \notin F^*_{mn}(u, v, r, c)$.

(ii) Let $A \in \cup F^*_{mn} \setminus \cap F^*_{mn}$. By (i), A has at least one zero entry. Since A belongs to some $F^*_{mn}(u, v, r, c)$, where u, v, r, c are strictly positive vectors, it follows that A has neither a zero row nor a zero column. Conversely, if A has a zero entry but no zero row or zero column, then by (i), $A \notin \cap F^*_{mn}$. Moreover, $A \in F_{mn}(e_n, e_m, r, c)$, where r and c are, respectively, the strictly positive vectors of row sums and column sums of A.

(iii) This equivalence follows directly from (i) and (ii). □

The next theorem shows that the classifications with respect to S^*_{mn} and F^*_{mn} coincide.

THEOREM 4.3. *We have* $\cap S^*_{mn} = \cap F^*_{mn}$ *and* $\cup S^*_{mn} = \cup F^*_{mn}$.

Proof. Trivially, $\cap F^*_{mn} \subseteq \cap S^*_{mn}$ and $\cup S^*_{mn} \subseteq \cup F^*_{mn}$. The reverse inclusions follow immediately from Corollary 3.5. □

Recall that a square matrix is said to be *completely reducible* if for some permutation matrix P, the matrix PAP^{T} is a direct sum of irreducible matrices.

THEOREM 4.4. *Let* $A \in R^{\mathbf{nn}}_{+0}$. *Then we have the following*:

(i) $A \in \cap E^*_{nn}$ *if and only if* A *is completely reducible and the diagonal elements of* A *are positive*.

(ii) $A \in \cup E^*_{nn} \setminus \cap E^*_{nn}$ *if and only if* A *is completely reducible,* $a_{ii} = 0$ *for some* $i \in \langle n \rangle$, *and* A *has no zero row or zero column*.

(iii) $A \notin \cup E^*_{nn}$ *if and only if either* A *is not completely reducible or* A *has a zero row or zero column*.

Proof. Since the conditions in (i)–(iii) are mutually exclusive and collectively exhaustive, it is enough to prove the "if" part in each of the three assertions.

(i) Suppose that A is completely reducible with positive diagonal elements. Let $u, v \in R^{\mathbf{n}}_{+}$ and let α and β be nonempty proper subsets of $\langle n \rangle$. Suppose that

$$A[\alpha \mid \beta] = 0. \tag{4.5}$$

Also, suppose that

$$A[\alpha' \mid \beta'] \neq 0. \tag{4.6}$$

Since A has positive diagonal elements it follows from (4.5) that $\alpha \cap \beta = \varnothing$, i.e., $\alpha \subseteq \beta'$. We claim that α is a proper subset of β'. Suppose to the contrary that $\alpha = \beta'$. Then (4.5) and (4.6) imply that $A[\beta' \mid \beta] = 0$ and $A[\beta \mid \beta'] \neq 0$, contradicting the assumption that A is completely reducible. Thus, $\gamma = \alpha \cup \beta$ is a proper subset of $\langle n \rangle$ and, since $\alpha \cap \beta = \varnothing$, we have

$$v^{\mathrm{T}}_{\alpha} u_{\alpha} + v^{\mathrm{T}}_{\beta} u_{\beta} = v^{\mathrm{T}}_{\gamma} u_{\gamma} < v^{\mathrm{T}} u, \tag{4.7}$$

implying that

$$v^{\mathrm{T}}_{\alpha} u_{\alpha} < v^{\mathrm{T}} u - v^{\mathrm{T}}_{\beta} u_{\beta} = v^{\mathrm{T}}_{\beta'} u_{\beta'}. \tag{4.8}$$

Now suppose that (4.5) holds and that

$$A[\alpha' \mid \beta'] = 0. \tag{4.9}$$

As before, (4.5) implies that $\alpha \subseteq \beta'$. Similarly, (4.9) implies that $\alpha' \subseteq \beta$, i.e., $\beta' \subseteq \alpha$. So, $\alpha = \beta'$, and hence $v^{\mathrm{T}}_{\alpha} u_{\alpha} = v^{\mathrm{T}}_{\beta'} u_{\beta'}$. It now follows from Theorem 3.10 that

$$A \in F^*_{mn}(u, v, u, v) = E^*_{nn}(u, v).$$

(ii) Suppose that A is completely reducible, that $a_{ii} = 0$ for some $i \in \langle n \rangle$, and that A has no zero row or zero column. We choose $u, v \in R^{\mathbf{n}}_{+}$ with $v_i u_i = \frac{2}{3}$ and $v^{\mathrm{T}} u = 1$. Then for $\alpha = \beta = \{i\}$ we have

$$v^{\mathrm{T}}_{\alpha} u_{\alpha} = \tfrac{2}{3} > \tfrac{1}{3} = v^{\mathrm{T}} u - v_i u_i = v^{\mathrm{T}}_{\beta'} u_{\beta'}.$$

Therefore, by Theorem 3.10 and Notation 2.10 we have $A \notin E^*_{nn}(u, v)$. We now have to show that $A \in \cup E^*_{nn}$. Since A is completely reducible, it follows that A is a direct sum of irreducible matrices. Furthermore, since A has no zero row or zero column, each of these irreducible matrices is nonzero and thus has a positive spectral radius. It now follows that we can find a matrix B which is positively diagonally equivalent to A, where

B is a direct sum of irreducible matrices with spectral radii 1. By the Perron–Frobenius theory for nonnegative matrices it follows that for some $u, v \in R_+^n$ we have $Bu = u$ and $v^T = v^T B$, i.e., $B \in E_{nn}(u, v)$. Hence $A \in E_{nn}^*(u, v) \subseteq \cup E_{nn}^*$.

(iii) In the case that A has a zero row or zero column the assertion is clear. Suppose that A is not completely reducible. Then there exist nonempty subsets α and β of $\langle n \rangle$ with $\alpha = \beta'$, such that $A[\alpha|\beta] = 0$ and $A[\alpha'|\beta'] \neq 0$. Since $\alpha = \beta'$, for every $u, v \in R_+^n$ we have $v_\alpha^T u_\alpha = v_{\beta'}^T u_{\beta'}$. By Theorem 3.10 it now follows that $A \notin E_{nn}^*(u,v)$. □

Our final observation shows that the requirement $u = v$ in $E_{nn}^*(u, v)$ or $r = c$ in $S_{mn}^*(r, c)$ does not yield new classifications. Specifically, let

$$\cap E_n^* = \bigcap_{u \in R_+^n} E_{nn}^*(u,u),$$

$$\cup E_n^* = \bigcup_{u \in R_+^n} E_{nn}^*(u,u),$$

$$\cap S_n^* = \bigcap_{r \in R_+^n} S_{nn}^*(r,r),$$

$$\cup S_n^* = \bigcup_{r \in R_+^n} S_{nn}^*(r,r).$$

THEOREM 4.10. *We have*

$$\cap S_n^* = \cap E_n^* = \cap E_{nn}^*,$$

$$\cup S_n^* = \cup E_n^* = \cup E_{nn}^*.$$

Proof. For $u \in R_+^n$, let $u^{(1/2)}$ be the vector in R_+^n with $(u^{(1/2)})_i = (u_i)^{1/2}$, $i = 1, \cdots, n$. Then Corollary 3.6 shows that $S_{nn}^*(u, u) = E_{nn}^*(u^{(1/2)}, u^{(1/2)})$, implying that $\cap E_n^* \subseteq \cap S_n^*$ and $\cup S_n^* \subseteq \cup E_n^*$. Next, the inclusions $\cap E_{nn}^* \subseteq \cap E_n^*$ and $\cup E_n^* \subseteq \cup E_{nn}^*$ are immediate, and the inclusions $\cap S_n^* \subseteq \cap E_{nn}^*$ and $\cup E_{nn}^* \subseteq \cup S_n^*$ follow directly from Corollary 3.6. Thus, the conclusions of our theorem have been established. □

REFERENCES

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

[2] S. FRIEDLAND AND S. KARLIN, *Some inequalities for the spectral radius of non-negative matrices and applications*, Duke Math. J., 62 (1975), pp. 459–490.

[3] M. V. MENON AND H. SCHNEIDER, *The spectrum of a nonlinear operator associated with a matrix*, Linear Algebra Appl., 2 (1969), pp. 321–334.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 4, October 1988

LINEAR PRESERVERS OF THE CLASS OF HERMITIAN MATRICES WITH BALANCED INERTIA*

STEPHEN PIERCE† AND LEIBA RODMAN‡

Abstract. Let $H(n)$ be the n^2-dimensional real vector space of Hermitian matrices. Assume n is even and greater than or equal to 4. Let T be an invertible linear transformation on $H(n)$ that maps the class of invertible, balanced inertia (signature zero) Hermitian matrices into itself. Then for some real number $c \neq 0$, and an invertible matrix S, $T(A) = cS^*AS$ or $T(A) = cS^*A^TS$, for all $A \in H(n)$. T is also classified in the case where $n = 2$.

Key words. linear preservers, inertia, Grassmannian

AMS(MOS) subject classifications. 15A04, 15A57

1. Introduction. Let $H(n)$ be the set of all $n \times n$ Hermitian matrices considered as a vector space of dimension n^2 over the real numbers $\mathbf{R}$. For nonnegative integers r, s, t such that $r + s + t = n$, we let (r, s, t) represent the inertia class of all A in $H(n)$ such that A has r positive, s negative, and t zero eigenvalues.

We are particularly interested in the nonsingular balanced inertia class $(k, k, 0)$ (so that $n = 2k$ is necessarily even). Suppose T is an invertible linear transformation mapping the class $(k, k, 0)$ into itself. The purpose of this paper is to classify all such T.

In previous related work [HR], [JP1], [JP2], the authors classified all invertible linear transformations T mapping the inertia class (r, s, t) into itself with the following exceptions: (i) the positive definite and negative definite inertia classes $(n, 0, 0)$ and $(0, n, 0)$; (ii) the nonsingular balanced inertia class $(k, k, 0)$ with $n = 2k$. In some cases, the assumption of invertibility of T could be removed without disturbing the result. We suspect this is always the case for nonsingular indefinite inertia classes when $n \geqq 2$. The problem is significantly more complicated for the definite inertia classes. For a discussion of results in this area see, for example, [C].

The following theorem is the main result of this paper. We assume $n = 2k$.

THEOREM 1.1. *Let T be an invertible linear transformation on $H(n)$, where $n \geqq 4$, and assume that $T(k, k, 0) \subset (k, k, 0)$. Then T is one of the following four types (here S stands for a fixed invertible $n \times n$ matrix)*: (1) $T(A) = S^*AS$; (2) $T(A) = -S^*AS$; (3) $T(A) = S^*A^TS$; (4) $T(A) = -S^*A^TS$.

Note that Theorem 1.1 fails for $n = 2$. Indeed, for a fixed $\alpha \in \mathbf{C}$, let T_α be the linear transformation on $H(2)$ that multiplies a_{12} by α and a_{21} by $\bar{\alpha}$. It is easy to see that if $|\alpha| > 1$, then T is invertible and preserves the inertia class $(1, 1, 0)$, but is not one of the four forms satisfying Theorem 1.1.

As a byproduct we also obtain a characterization of the invertible linear transformations on $H(n)$ that preserve the set of matrices with a k-dimensional isotropic subspace.

THEOREM 1.2. *Let T: $H(n) \to H(n)$ be an invertible linear transformation, where n is even and greater than or equal to 4. Suppose that, for any $A \in H(n)$ such that there is a k-dimensional subspace $V \subset \mathbf{C}^n$ with $x^*Ay = 0$ for all $x, y \in V$, the matrix $T(A)$ also enjoys this property (as before, $k = n/2$). Then T is one of the four types described in Theorem* 1.1.

* Received by the editors July 30, 1987; accepted for publication (in revised form) February 5, 1988.

† Department of Mathematical Sciences, San Diego State University, San Diego, California 92182. The research of this author was partially supported by National Science Foundation grant DMS-0861959.

‡ Department of Mathematics, Arizona State University, Tempe, Arizona 85287. The research of this author was partially supported by National Science Foundation grant DMS-8501794.

Indeed, the class of $A \in H(n)$ such that $x^*Ay = 0$ for all $x, y \in V$, for some k-dimensional subspace $V \subset \mathbf{C}^n$, coincides with the closure $\overline{(k, k, 0)}$ of $(k, k, 0)$. It is easy to see (arguing by contradiction) that $T\overline{(k, k, 0)} \subset \overline{(k, k, 0)}$ implies $T(k, k, 0) \subset (k, k, 0)$. Indeed, assume $A \in T(k, k, 0)$, where the invertible linear transformation T maps $\overline{(k, k, 0)}$ into itself. Then A belongs to $\overline{(k, k, 0)}$. On the other hand, $(k, k, 0)$ is an open set and T is an open map. Thus A belongs to the open set $T(k, k, 0)$, which is contained in $\overline{(k, k, 0)}$. As the interior of $\overline{(k, k, 0)}$ coincides with $(k, k, 0)$, the matrix A actually belongs to $(k, k, 0)$. Now use Theorem 1.1.

The methods of this work are almost completely independent of [HR], [JP1], and [JP2], although we do rely on one particular lemma in [JP2]. We depend heavily on an examination of the Grassmannian $G(n)$ consisting of the subspaces of $\mathbf{C}^n$.

In the case where $n = 2$, we have an entirely different problem. The conclusion of Theorem 1.1 is no longer true as noted above. To state Theorem 1.3, we first define for $r, s \in \mathbf{R}$, a linear map $D_{r,s}$ on $H(2)$ by

$$D_{r,s}: \begin{bmatrix} a & u+iv \\ u-iv & b \end{bmatrix} \to \begin{bmatrix} a & ru+siv \\ ru-siv & b \end{bmatrix}.$$

As long as $|r|$ and $|s|$ are greater than or equal to 1, $D_{r,s}$ preserves $K(2)$, the set of indefinite 2×2 Hermitian matrices. If $r = s$, we will just write D_r.

Finally we note that the linear preservers of $K(2)$ preserve the negative cone of the quadratic form $ab - u^2 - v^2$ obtained from the Hermitian matrix given above in the definition of $D_{r,s}$. This quadratic form has signature (1, 3, 0), precisely as in the form appearing in special relativity theory.

THEOREM 1.3. *Let T be an invertible linear map on $H(2)$, which maps $K(2)$ into itself. Then T is a product of maps of the type given in Theorem* 1.1 *and maps of the form $D_{r,s}$ with $|r|, |s| \geqq 1$.*

2. The Grassmannian. In this section, we note some properties of the Grassmannian $G(n)$ consisting of all subspaces of $\mathbf{C}^n$. We endow $\mathbf{C}^n$ with the standard inner product $\langle , \rangle$ and the corresponding norm $\|x\|^2 = |x_1|^2 + \cdots + |x_n|^2$.

For V and W in $G(n)$, we introduce the *gap*

$$\theta(V, W) = \|P_V - P_W\|,$$

where P_V is the orthogonal projector on V and $\|\cdot\|$ is the operator norm, i.e., $\|A\|$ is the maximum singular value of A. Now the gap θ satisfies all the properties of a metric, thus turning $G(n)$ into a metric space.

PROPOSITION 2.1. *With the metric θ, $G(n)$ is a compact (and hence complete) metric space. Moreover, $G(n)$ has precisely $n + 1$ connected components, each connected component being the set $G_k(n)$ of all subspaces in $\mathbf{C}^n$ of a fixed dimension k, $0 \leqq k \leqq n$.*

PROPOSITION 2.2. *Let $V_m \in G(n)$, $m = 1, 2, \cdots$, and assume that*

$$\lim_{m \to \infty} \theta(V, V_m) = 0$$

for some $V \in G(n)$. Then V consists of precisely those vectors $x \in \mathbf{C}^n$ for which there is a sequence $\{x_m\}$, $m = 1, 2, \cdots$, where $x_m \in \mathbf{C}^n$ and $\lim_{m \to \infty} x_m = x$ and $x_m \in V_m$ for $m = 1, 2, \cdots$.

These two propositions are well known and appear in, e.g., [GLR1] or [GLR2].

3. Inertia classes. Recall that we are interested in the balanced inertia class $(k, k, 0)$, with $n = 2k$, $k > 1$. First we note that the closure of $(k, k, 0)$, $\overline{(k, k, 0)}$ is the union of all inertia classes with positive and negative inertia not exceeding k. Alternatively,

$A \in (\overline{k,k,0})$ if and only if there is a k-dimensional A-isotropic subspace, i.e., a subspace $V \subset \mathbf{C}^n$ such that $\langle Ax, y\rangle = 0$ for all $x, y \in V$.

PROPOSITION 3.1. *Let $A, B \in (\overline{k,k,0})$, with A invertible. Then A and B have a common isotropic k-dimensional subspace if and only if*

$$\lambda A + \mu B \in (\overline{k,k,0}) \tag{3.1}$$

for all $\lambda, \mu \in \mathbf{R}$.

Proof. The result is obvious if A and B have a common k-dimensional isotropic subspace. Thus suppose (3.1) holds. Without loss of generality, we assume that the pair A, B is in the canonical pair form with respect to simultaneous congruence (see [HJ] or [T]). Briefly, denote by $J_p(\alpha)$ the $p \times p$ Jordan block with eigenvalue α. Then A and B are simultaneously direct sums of equal-sized blocks of the form $\pm P_r$ and $P_r K$, respectively, where P_r is the $r \times r$ permutation matrix satisfying $p_{ij} = 1$ if $i + j = r + 1$; $p_{ij} = 0$ otherwise, and K is either $J_r(\alpha)$ for some real α or $J_{r/2}(\alpha) \oplus J_{r/2}(\bar{\alpha})$ for some nonreal α. It is easy to see that, in the cases when $K = J_r(\alpha)$ with real α and even r or $K = J_{r/2} \oplus J_{r/2}(\bar{\alpha})$ with nonreal α, the matrices $\pm P_r$ and $P_r K$ have a common $r/2$-dimensional isotropic subspace. Thus we need only consider the cases when $K = J_r(\alpha)$ with real α and odd r. As $A \in (k, k, 0)$, it follows that the number of blocks $-P_r$ in A with odd r is equal to the number of blocks P_r with odd r. This observation allows us to reduce the consideration to the case when

$$\begin{aligned} A &= P_{r_1} \oplus \cdots \oplus P_{r_q} \oplus (-P_{s_1}) \oplus \cdots \oplus (-P_{s_q}), \\ B &= P_{r_1} J_{r_1}(\alpha_1) \oplus \cdots \oplus P_{r_q} J_{r_q}(\alpha_q) \oplus P_{s_1} J_{s_1}(\beta_1) \oplus \cdots \oplus P_{s_q} J_{s_q}(\beta_q), \end{aligned} \tag{3.2}$$

where $r_1, \cdots, r_q, s_1, \cdots, s_q$ are odd and $\alpha_1, \cdots, \alpha_q, \beta_1, \cdots, \beta_q$ are real (without loss of generality assume $\alpha_1 \leqq \cdots \leqq \alpha_q$; $\beta_1 \geqq \cdots \geqq \beta_q$). As we can easily see, condition (3.1) for A and B given by (3.2) is equivalent to the fact that the number of positive numbers in the list $\{\lambda + \alpha_1, \cdots, \lambda + \alpha_q, -\lambda + \beta_1, \cdots, -\lambda + \beta_q\}$ is precisely q for every real λ different from the $-\alpha_j$'s and β_j's. This can happen only if $\alpha_j = -\beta_j$ for $j = 1, \cdots, q$. (Indeed, letting $\lambda = -\alpha_1 + \varepsilon$ with small $\varepsilon > 0$ and $\lambda = \beta_1 - \varepsilon$ with small $\varepsilon > 0$, we conclude that $\alpha_1 + \beta_1 = 0$. Now apply induction on q.) Now the existence of a common isotropic k-dimensional subspace for A and B is evident (assuming $q = 1$ for the simplicity of notation. Letting $r = r_1$, $s = s_1$, we see that one such subspace is spanned by the vectors $e_1, \cdots, e_{(r-1)/2}, e_{r+1}, \cdots, e_{r+(s-1)/2}, e_{(r+1)/2} + e_{r+(s+1)/2}$, where e_j is the jth standard vector). □

For related results, see [JR], [RU], or [RR].

Now let T: $H(n) \to H(n)$ be an invertible linear transformation such that T maps $(k, k, 0)$ into $(k, k, 0)$. Then clearly $T(\overline{k,k,0}) \subset (\overline{k,k,0})$.

PROPOSITION 3.2. *Let $A \in (k, k, 0)$, $B \in (\overline{k,k,0})$, and assume that A and B have a common isotropic k-dimensional subspace. Then the same is true of $T(A)$ and $T(B)$.*

Proof. By Proposition 3.1, $\lambda A + \mu B \in (\overline{k,k,0})$ for all real λ, μ. Thus, $\lambda T(A) + \mu T(B) = T(\lambda A + \mu B) \in (\overline{k,k,0})$, and the proposition follows from Proposition 3.1. □

For a given $A \in (\overline{k,k,0})$, let $Z(A) \subset G_k(n)$ be the set of all k-dimensional A-isotropic subspaces. Proposition 2.2 shows that $Z(A)$ is closed in $G_k(n)$.

For a given $A \in (\overline{k,k,0})$, let

$$D(A) = \{B \in (\overline{k,k,0}) \mid Z(A) \cap Z(B) \neq \varnothing\}.$$

PROPOSITION 3.3. *If T is as above, then $T(D(A)) \subset D(T(A))$.*

Proof. Assume first that $A \in (k, k, 0)$. Let $B \in D(A)$. By Proposition 3.2, $T(B) \in D(T(A))$, and we are done. Next, suppose that $A \in (\overline{k,k,0})$ is singular. Let $B \in D(A)$ and let V be a common isotropic k-dimensional subspace for A and B. Find a sequence $A_m \in (k, k, 0)$ such that $V \in Z(A_m)$ for all m and $A_m \to A$ as $m \to \infty$.

For example, we may assume that

$$A = \begin{bmatrix} 0 & A_1 \\ A_1^* & A_2 \end{bmatrix}, \qquad A_m = \begin{bmatrix} 0 & A_{1m} \\ A_{1m}^* & A_{2m} \end{bmatrix},$$

where A_{1m}, A_{2m} are $k \times k$ with invertible A_{1m} and $\lim_{m\to\infty} A_{im} = A_i$, $i = 1, 2$.

Obviously, $B \in D(A_m)$ for all m. By the part of the proposition already proved, $T(B) \in D(T(A_m))$ for all m. Thus there is a common k-dimensional isotropic subspace V_m for $T(B)$ and $T(A_m)$. By Proposition 2.1, choose a converging subsequence $V_{m_p} \to V$ for some V. As $T(A_m) \to T(A)$, Proposition 2.2 shows that V is isotropic for $T(B)$ and $T(A)$. Since $\dim V_{m_p} = k$, we also have $\dim V = k$. Hence $T(B) \in D(T(A))$. □

4. More about the Grassmannian. With $2k = n$, consider the set $G_k(n)$ of all k-dimensional subspaces of $\mathbf{C}^n$, topologized by the gap metric. We introduce the standard structure of a real analytic manifold on $G_k(n)$.

Let α be a selection of k indices $\alpha_1 < \cdots < \alpha_k$ from $\{1, \cdots, n\}$. Let V_α be a subset (*chart*) of $G_k(n)$ defined as follows: $V \in V_\alpha$ if and only if V is the column space of some $n \times k$ complex matrix X_V whose k rows with indices α form the $k \times k$ identity matrix. Note that X_V is uniquely defined by V. The set V_α is open and dense in $G_k(n)$. Define the map

$$f_\alpha : V_\alpha \to \mathbf{R}^{2k^2}$$

by the property that $f_\alpha(V)$ is the $k \times k$ matrix formed by the rows of X_V other than the rows α. Clearly f_α is bijective and maps V_α homeomorphically onto R^{2k^2}. Observe that, for any two selections α and β, the set $f_\alpha(V_\alpha \cap V_\beta)$ is open in R^{2k^2}; in fact, the complement of $f_\alpha(V_\alpha \cap V_\beta)$ is a real algebraic set. Also observe that the map

$$f_\beta f_\alpha^{-1} : f_\alpha(V_\alpha \cap V_\beta) \to R^{2k^2} \tag{4.1}$$

is real analytic in the $2k^2$ real variables representing a point in R^{2k^2}. Note that $G_k(n)$ is the union of V_α for all selections α.

A set $S \subset G_k(n)$ will be called an *analytic set* if the following holds: For every $V \in G_k(n)$ and a chart V_α such that $V \in V_\alpha$, there exist an open neighborhood W of $f_\alpha(V)$ and a real analytic function g: $W \to \mathbf{R}$ such that

$$X \in f_\alpha^{-1}(W) \cap S$$

if and only if $g(f_\alpha(X)) = 0$. Since (4.1) is real analytic this definition does not depend on the choice of the chart V_α. Clearly, an analytic set is closed (in the gap metric). Finite intersections and unions of analytic sets are again analytic.

PROPOSITION 4.1. *Let $A \in (k, k, 0)$. Then $Z(A)$ is an analytic set.*

Proof. As $Z(A)$ is a closed set in the gap metric, we must check the property that appears in the definition of an analytic set only for $V \in Z(A)$. Let V_α be a chart such that $V \in V_\alpha$ and for notational simplicity suppose that $\alpha = (1, \cdots, k)$. Then V is the range (i.e., the column space) of

$$\begin{bmatrix} I_k \\ X_0 \end{bmatrix}$$

for some $k \times k$ matrix X_0. Write

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^* & A_{22} \end{bmatrix};$$

then the range of $[I_k \; X]^T \in Z(A)$ if and only if the Hermitian form

$$[I \quad X^*]A\begin{bmatrix} I \\ X \end{bmatrix} = 0. \tag{4.2}$$

Write the (p, q) entry of X as $x_{pq} + iy_{pq}$, $p, q = 1, \cdots, k$. Then the system of equations (4.2) can be expressed in the form

$$g_s(x_{pq}, y_{pq}) = 0, \qquad s = 1, \cdots, s_0,$$

where $g_s(x_{pq}, y_{pq})$ is a polynomial whose coefficients are real polynomials in the entries of A. Now set

$$g(x_{pq}, y_{pq}) = \sum_{s=1}^{s_0} g_s(x_{pq}, y_{pq})^2$$

to satisfy the definition of an analytic set. □

A continuous map $T: G_k(n) \to G_k(n)$ will be called *analytic* if for any two charts V_α, V_β the map

$$f_\beta T f_\alpha^{-1}: f_\alpha(X) \to R^{2k^2}$$

is analytic, where $X = \{V \in V_\alpha \mid TV \in V_\beta\}$. (The continuity of T ensures that X, and hence $f_\alpha(X)$, are open sets; thus the analyticity of the map $f_\beta T f_\alpha^{-1}$ defined on $f_\alpha(X)$ makes sense.)

PROPOSITION 4.2. *Let T be an $n \times n$ invertible matrix. Then the map $\tilde{T}: G_k(n) \to G_k(n)$ defined by $\tilde{T}(V) = \{Tx \mid x \in V\}$ is analytic.*

Analogously (using the charts) we define the analyticity of a map $T: U \to R^p$, where U is an open set in $G_k(n)$, and of a map $T: R^p \to G_k(n)$ (in all these cases continuity of T is a prerequisite for analyticity).

A closed set $S \subset G_k(n)$ is called a *real analytic manifold* if for every $V \in S$ and a chart V_α such that $V \in V_\alpha$ there is an open neighborhood $\mathcal{U}$ of U such that $\mathcal{U} \subset V_\alpha$ and real analytic functions $g_1, \cdots, g_{2k^2}$ on $f_\alpha(\mathcal{U})$ exist with the following properties:

$$\det\left[\frac{\partial g_i(t)}{\partial t_j}\right]_{i,j=1}^{2k^2} \neq 0, \qquad t = (t_1, \cdots, t_{2k^2}) \in f_\alpha(\mathcal{U}),$$

$$f_\alpha(\mathcal{U} \cap S) = \{t \in f_\alpha(\mathcal{U}) \mid g_{p+1}(t) = \cdots = g_{2k^2}(t) = 0\},$$

where p is some integer. It easily follows that the number p does not depend on the choice of $g_1, \cdots, g_{2k^2}$ (subject to the properties mentioned above) and that p is constant for every V belonging to a fixed connected component of S. The maximum of the numbers p will be called the *dimension* of S.

It is a standard fact (see, e.g., [GN] or [W]) that any analytic set S can be represented as the union of a finite number of real analytic manifolds. The maximal dimension of these analytic manifolds will be called the *dimension* of S.

Let $S \in G_k(n)$. Define

$$\Gamma(S) = \{A \in H(n) \mid \text{there is a subspace } V \in S \text{ which is } A\text{-isotropic}\}.$$

The set $\Gamma(S)$ can be alternatively described as follows. Consider $H(n) \times G_k(n)$, which is a real analytic manifold, and the set U of all pairs $(A, V) \in M_n \times G_k(n)$ such that $V \in S$ and V is A-isotropic. It is easy to see that U is an analytic manifold; then $\Gamma(S)$ is the projection of U to the first component.

We can show that $\Gamma(S)$ is a semianalytic set (see, e.g., [L]). That is, locally $\Gamma(S)$ is given as the set of solutions of a system of inequalities $f_1(A) \geqq 0, \cdots, f_s(A) \geqq 0$, where the $f_j(A)$ are real analytic functions of the entries of $A \in H_n$. As any semianalytic set is locally the union of a finite number of real analytic manifolds, we can define the dimension of $\Gamma(S)$ as the maximal dimension of a real analytic manifold contained in $\Gamma(S)$.

THEOREM 4.3. *If* $S_1, S_2 \subset G_k(n)$ *are analytic sets and* $\dim S_1 < \dim S_2$, *then* $\dim \Gamma(S_1) < \dim \Gamma(S_2)$.

Proof. Let $V_1 \in S_1$, $V_2 \in S_2$, where V_2 is such that the dimension of S_2 coincides with the dimension of S_2 in a neighborhood of V_2. It is sufficient to prove that the dimension of $\Gamma(W_2 \cap S_2)$ is bigger than the dimension of $\Gamma(W_1 \cap S_1)$, where W_1 and W_2 are small neighborhoods of V_1 and V_2, respectively. We can assume also that S_1 and S_2 are analytic manifolds. This follows from the fact that any analytic set in $G_k(n)$ is a finite union of analytic manifolds. Applying an invertible linear transformation on $\mathbf{C}^n$ which maps V_1 onto V_2, we can assume that $V_1 = V_2 = V$, and set $W_1 = W_2 = W$. Without loss of generality, assume that V is the column space of $[I_k \; A_0]^T$, where A_0 is $k \times k$ invertible. Thus

$$S_1 = \left\{ \text{range} \begin{bmatrix} I \\ X(t) \end{bmatrix} \middle| \begin{array}{l} X(t) \text{ depends analytically on a parameter } t \text{ in } U_1, \\ \qquad\qquad \text{where } U_1 \subset R^{q_1} \text{ is an open neighborhood of zero} \end{array} \right\},$$

$$S_2 = \left\{ \text{range} \begin{bmatrix} I \\ Y(s) \end{bmatrix} \middle| \begin{array}{l} Y(s) \text{ depends analytically on a parameter } s \text{ in } U_2, \\ \qquad\qquad \text{where } U_2 \subset R^{q_2} \text{ is an open neighborhood of zero} \end{array} \right\}.$$

Also $X(0) = A_0$; $Y(0) = A_0$; $q_1 < q_2$ and there are analytic functions $f_1 : X(U_1) \to U_1$, $f_2 : Y(U_2) \to U_2$ such that $f_1(X(t)) = t$ for all $t \in U_1$, and $f_2(Y(s)) = s$ for all $s \in U_2$. We may assume that

$$U_1 = \{(x_1, \cdots, x_{q_1}, \cdots, 0)^T \in R^{q_2} \mid -\varepsilon < x_i < \varepsilon\},$$

$$U_2 = \{(x_1, \cdots, x_{q_2})^T \in R^{q_2} \mid -\varepsilon < x_i < \varepsilon\},$$

and $\varepsilon > 0$ is suitably small, so that $U_1 \subset U_2$. Since A_0 is assumed to be invertible, we can assume that $Y(s)$ is invertible for all $s \in U_2$, and $X(t)$ is invertible for all $t \in U_1$.

Applying to S_1 the invertible linear transformation (which is a real analytic function of $t \in U_1$)

$$\begin{bmatrix} I & 0 \\ 0 & Y(t)X(t^{-1}) \end{bmatrix}, \qquad t \in U_1,$$

we can assume that actually $X(t) = Y(t)$ for all $t \in U_1$. (This follows from the equality $\Gamma(X(S)) = \{X^*AX \mid A \in \Gamma(S)\}$, where X is invertible.)

Let $H(k)$ be the vector space of all $k \times k$ Hermitian matrices, and let $F(k)$ be the vector space (over $\mathbf{R}$) of all pairs of matrices (Z_1, Z_2), where Z_1 is $k \times k$ Hermitian,

and Z_2 is any complex $k \times k$ matrix. For any $Y(s)$, $s \in U_2$, let $\tilde{Y}(s)$ be the linear transformation from $F(k)$ to $H(k)$ defined by

$$\tilde{Y}(s)(Z_1, Z_2) = -Y(s)^* Z_1 Y(s) - Y(s)^* Z_2^* - Z_2 Y(s).$$

Observe that

$$\Gamma\left(\text{Range}\begin{bmatrix} I \\ Y(s) \end{bmatrix}\right)$$

can be identified with the graph of the linear transformation $\tilde{Y}(s)$:

$$\Gamma\left(\text{range}\begin{bmatrix} I \\ Y(s) \end{bmatrix}\right) = \{(Z, \tilde{Y}(s)Z) \mid Z \in F(k)\}. \tag{4.3}$$

In this identification,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^* & A_{22} \end{bmatrix}$$

from the left-hand side of (4.3) is identified with $\{(A_{22}, A_{12}), A_{11}\}$ on the right-hand side of (4.3).

Define the transformation

$$\tilde{F}: U_2 \to L(F(k), H(k)),$$

where $L(F(k), H(k))$ stands for the set of all linear transformations from $F(k)$ to $H(k)$ as follows:

$$F(s) = \tilde{Y}(s), \qquad s \in U_2.$$

Then obviously $\tilde{F}$ is real analytic. According to (4.3), the theorem will be proved if we can show that $\tilde{F}$ has a real analytic inverse $\tilde{F}^{-1}$: $\tilde{F}(U_3) \to U_3$ for some neighborhood U_3 of zero in R^{q_2}. By the implicit function theorem, it is sufficient to verify that

$$\left.\frac{\partial \tilde{F}}{\partial x_j}\right|_{s=0}, \qquad j = 1, \cdots, q_2$$

are linearly independent (here $x_1, \cdots, x_{q_2}$ are the coordinates of $s \in R^{q_2}$). Computation shows that

$$\left.\frac{\partial \tilde{F}}{\partial x_j}\right|_{s=0}(Z_1, Z_2) = -\left[\frac{\partial Y(0)^*}{\partial x_j} Z_1 Y(0) + Y(0)^* Z_1 \frac{\partial Y(0)}{\partial x_j} + \frac{\partial Y(0)^*}{\partial x_j} Z_2^* + Z_2 \frac{\partial Y(0)}{\partial x_j}\right].$$

So, if some linear combination of

$$\left.\frac{\partial \tilde{F}}{\partial x_j}\right|_{s=0}$$

is zero, say

$$\sum_{j=1}^{q_2} \alpha_j \left.\frac{\partial \tilde{F}}{\partial x_j}\right|_{s=0} = 0, \qquad \alpha_j \in \mathbf{R},$$

then, denoting

$$U = \sum_{j=1}^{q_2} \alpha_j \frac{\partial Y(0)}{\partial x_j},$$

we have

$$U^*(Z_1Y(0)+Z_2^*)+(Y(0)^*Z_1+Z_2)U=0$$

for all Z_1 and Z_2. Choosing Z_1 and Z_2 so that $Z_1Y(0) + Z_2^* = I$, and choosing Z_1 and Z_2 again so that $Z_1Y(0) + Z_2^* = iI$, we obtain $U = 0$. As $\partial Y(0)/\partial x_j$ are linearly independent, $\alpha_j = 0$. □

COROLLARY 4.4. *Let* $T: H(n) \to H(n)$ *be a linear invertible transformation such that* $T(\overline{k,k,0}) \subset (\overline{k,k,0})$. *Then for every* $A \in (\overline{k,k,0})$ *we have*

$$\dim Z(T(A)) \geqq \dim Z(A).$$

Proof. Suppose not; then dim $Z(A) >$ dim $Z(T(A))$ for some $A \in (\overline{k,k,0})$. By Theorem 4.3 we have dim $D(A) >$ dim $D(T(A))$. But dim $D(A) =$ dim $D(T(A))$, since an invertible linear transformation preserves the dimension of semianalytic sets. Thus dim $D(A) >$ dim $D(T(A))$ contradicts Proposition 3.3. □

5. Deductions from Corollary 4.4. To make use of Corollary 4.4 we need to compute dim $Z(A)$ for certain Hermitian matrices A.

LEMMA 5.1. *If* A *is an* $n \times n$ *rank* 1 *Hermitian matrix, then the dimension of* $Z(A)$ *is* $2k^2 - 2k$, *where, as usual,* $2k = n$.

Proof. It is easily seen that $V \in Z(A)$ if and only if dim $V = k$ and $V \subset \ker(A)$. Thus $Z(A)$ can be identified with the set of all k-dimensional subspaces in a $(2k-1)$-dimensional complex space. The real dimension of this set is $2k(k-1) = 2k^2 - 2k$. □

Recall that we always count *real* dimension.

THEOREM 5.2. *Let* $k \geqq 2$ *and* $A \in (1, 1, n-2)$. *Then* dim $Z(A) = 2k^2 - 2k + 1$.

Proof. Without loss of generality we may assume that A is diagonal with precisely one 1, precisely one -1, and $n-2$ zeros on the main diagonal. We shall consider only the chart U_{α_0}, where $\alpha_0 = (1, \cdots, k)$. Then $Z(A) \cap U_{\alpha_0}$ can be identified with the solutions X of the following equation:

$$[I \quad X^*](-A_1 \oplus A_2)\begin{bmatrix} I \\ X \end{bmatrix} = 0, \tag{5.1}$$

where all the matrices in (5.1) have been partitioned into $k \times k$ blocks, $A_1 = I_{p_1} \oplus -I_{q_1} \oplus 0_{r_1}$, $A_2 = I_{p_2} \oplus -I_{q_2} \oplus 0_{r_2}$, $p_j + q_j + r_j = n, j = 1, 2$, and $p_2 + q_1 = p_1 + q_2 = 1$. Rewrite (5.1) in the form

$$X^*A_2X = A_1 \tag{5.2}$$

and observe that the necessary condition for solvability of (5.2) is that $p_2 \geqq p_1$ and $q_2 \geqq q_2$. Three cases can occur:

(1) $p_2 = q_2 = 1, p_1 = q_1 = 0$,
(2) $p_2 = p_1 = 1, q_1 = q_2 = 0$,
(3) $p_2 = p_1 = 0, q_1 = q_2 = 1$.

Consider case (1). Equation (5.1) takes the form

$$X^* \operatorname{diag}(1, -1, 0, \cdots, 0)X = 0. \tag{5.3}$$

Write $X = [x_{\alpha\beta}]$, $\alpha, \beta = 1, \cdots, k$ and set

$$\tilde{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}.$$

Let $X_1 = (x_{13}, \cdots, x_{1k})$ and $X_2 = (x_{23}, \cdots, x_{2k})$. Then (5.3) amounts to solving simultaneously the following equations:

$$\tilde{X}^* \operatorname{diag}(1,-1)\tilde{X} = 0, \tag{5.4}$$

$$\tilde{X}^* \begin{bmatrix} X_1 \\ -X_2 \end{bmatrix} = 0, \tag{5.5}$$

$$X_1^* X_1 = X_2^* X_2. \tag{5.6}$$

First we shall count the dimension of the set of solutions $\tilde{X}$ to (5.4). Clearly (5.4) is equivalent to

$$|x_{11}|^2 = |x_{21}|^2, \qquad |x_{12}|^2 = |x_{22}|^2,$$

$$\bar{x}_{11}x_{12} - \bar{x}_{21}x_{22} = 0.$$

First assume that $x_{11} \neq 0$. Then $x_{21} = e^{i\alpha}x_{22}$, $\alpha \in \mathbf{R}$, and $x_{12} = \bar{x}_{21}x_{12}/x_{11} = e^{-i\alpha}x_{22}$. Thus

$$\tilde{X} = \begin{bmatrix} x_{11} & e^{-i\alpha}x_{22} \\ e^{i\alpha} & x_{22} \end{bmatrix}$$

is given by five real parameters x_{11}, x_{22}, α (x_{11} and x_{22} being arbitrary complex numbers). Observe that

$$\ker \tilde{X}^* = \operatorname{span} \begin{bmatrix} 1 \\ -e^{i\alpha} \end{bmatrix}. \tag{5.7}$$

A similar argument shows that when $x_{22} \neq 0$, $\tilde{X}$ is given by five real parameters and $\ker \tilde{X}^*$ has the form (5.7). Finally, if $\tilde{X} \neq 0$, at least either $x_{11} \neq 0$ or $x_{22} \neq 0$. In view of (5.7), the solution $[X_1\ X_2]^T$ of (5.5) (where $\tilde{X}$ is considered as given) is given by $2(k-2)$ parameters in X_1; then

$$X_2 = e^{i\alpha}X_1. \tag{5.8}$$

If (5.8) holds, then (5.6) is satisfied automatically. The total number of parameters to describe the solution X of (5.3) is

$$5 + 2(k-2) + 2(k-2)k = 2k^2 - 2k + 1$$

(the term $2(k-2)k$ in the left-hand side describes the parameters in x_{ij}, $i \geqq 3$, $1 \leqq j \leqq k$, which are absolutely free).

Consider case (2). (Case (3) is obtained from case (2) by replacing A with $-A$ and will not be discussed.) Then

$$X^* \operatorname{diag}(1,0,\cdots,0)X = \operatorname{diag}(1,0,\cdots,0). \tag{5.9}$$

Write

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix},$$

where X_{11} is 1×1 and X_{22} is $(k-1) \times (k-1)$. Then (5.9) is equivalent to

$$|x_{11}|^2 = 1, \quad X_{12}^* X_{12} = 0, \quad \bar{X}_{11}X_{12} = 0.$$

Thus $X_{12} = 0$, and the solutions X of (5.9) are described by one real parameter of X_{11} and $2k(k-1)$ real parameters of X_{21} and X_{22}. The total number of parameters is $2k(k-1) + 1 = 2k(k-1)$. □

THEOREM 5.3. *Let* $A \in (k-1, k-1, 2)$. *Then* $\dim Z(A) \geq k^2 + 1$.

Proof. Let J be the $k \times k$ matrix $I_{k-1} \oplus 0$. Without loss of generality, we may assume that $A = J \oplus -J$. It is sufficient to prove that the set of solutions X of

$$[I \quad X^*]\begin{bmatrix} J & 0 \\ 0 & -J \end{bmatrix}\begin{bmatrix} I \\ X \end{bmatrix} = 0 \tag{5.10}$$

has dimension $k^2 + 1$. Write

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix},$$

where X_{11} is $(k-1) \times (k-1)$ and X_{22} is 1×1. It is easily seen that (5.10) holds if and only if X_{11} is unitary and $X_{12} = 0$. Since the set of $(k-1) \times (k-1)$ unitary matrices is $(k-1)^2$-dimensional, the total dimension of the set of solutions of (5.10) is $(k-1)^2 + 2k = k^2 + 1$ (the $2k$ dimensions appear because X_{21} and X_{22} are arbitrary). □

THEOREM 5.4. *If* $A \in (k, k, 0)$, *then* $\dim Z(A) = k^2$.

Proof. We can assume that A is a diagonal matrix with k 1's and $k-1$'s on the main diagonal. It suffices to consider the intersection of $Z(A)$ with one chart in the Grassmannian, say with the chart U_α, where $\alpha = (1, \cdots, k)$. In other words, we must compute the dimension of the set of solutions X of

$$[I \quad X^*]\begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}\begin{bmatrix} I \\ X \end{bmatrix} = 0, \tag{5.11}$$

where $A = J_1 \oplus J_2$, and J_1 and J_2 are $k \times k$ diagonal matrices with 1's and -1's on the main diagonal. Equation (5.11) is equivalent to

$$X^* J_2 X = -J_1. \tag{5.12}$$

Thus $-J_1$ and J_2 are congruent, and we may assume that $J_1 = -J_2$. Now X satisfies (5.12) if and only if X is an element of the J_2-unitary group. The (real) dimension of this group is k^2 (see, e.g., § X.2 in [H]), and we are done. □

COROLLARY 5.5. *Let* T: $H(n) \to H(n)$ *be an invertible linear transformation such* $T(k, k, 0) \subset (k, k, 0)$. *If* $A \in \overline{(k-1, k-1, 2)}$, *then* $T(A)$ *is singular.*

Proof. Since the set of singular Hermitian matrices is closed, it is sufficient to assume that $A \in (k-1, k-1, 2)$. Now $\dim Z(A) \geq k^2 + 1$ by Theorem 5.3, and by Corollary 4.4 we have $\dim Z(T(A)) \geq k^2 + 1$. In view of Theorem 5.4, the matrix $T(A)$ must be singular. □

THEOREM 5.6. *Let* T: $H(n) \to H(n)$ *be an invertible linear transformation such that* $T(k, k, 0) \subset (k, k, 0)$, *where* $k \geq 2$. *Then*

$$T(1, 1, n-2) \subset (1, 1, n-2). \tag{5.13}$$

Proof. Arguing by contradiction, we assume that $T(A)$ is not in $(1, 1, n-2)$ for some $A \in (1, 1, n-2)$. We cannot have rank $T(A)) = 1$ because of Lemma 5.1 and Theorem 5.2. Thus we may assume (replacing A by $-A$ if necessary) that $T(A)$ has at least two positive eigenvalues. Since A belongs to $\overline{(k-1, k-2, 2)}$, by Corollary 5.5 the matrix $T(A)$ is singular. Using Lemma 3 in [JP2], we choose a rank 1 Hermitian matrix B such that

$$i_+(T(A+B)) > i_+(T(A)), \qquad i_-(T(A+B)) \geq i_-(T(A)),$$

where i_+ and i_-, respectively, represent the number of positive and negative eigenvalues.

There are two possibilities:

(i) $n = 4 (k = 2)$. Clearly $A + B \in (\overline{2,2,0})$, but $i_+(T(A+B)) \geqq 3$. This contradicts the assumption that $T(\overline{2,2,0}) \subset (\overline{2,2,0})$.

(ii) $n \geqq 6$. Then $A + B \in (\overline{k-1, k-1, 2})$. Again, from Corollary 5.5, it follows that $T(A + B)$ is singular and $i_+(T(A + B)) \geqq 3$. Replacing $A + B$ with A, we repeat our procedure, obtaining a contradiction if $n = 6$ and iterating again if $n \geqq 8$. Eventually, of course, we will obtain a contradiction. □

It follows that T preserves the class $(1, 1, k - 2)$. It then follows from the main result in [JP2] that T has the required form, and hence Theorem 1.1 is established.

6. Proof of Theorem 1.3. Let $E_{i,j}$ be the matrix with 1 in position (i, j) and zero elsewhere. Suppose first that T maps every matrix in $\overline{K(2)}$ to a matrix in $K(2)$, i.e., every matrix of rank 1 is mapped to a matrix of negative determinant (such T do exist). Let S be the set of all $A \in \overline{K(2)}$ with Frobenius norm 1. Then S is compact and thus $T(S)$ is a compact subset of $K(2)$. It follows that there is an ε, $0 < \varepsilon < 1$, such that $D_\varepsilon T$ still maps $K(2)$ into $K(2)$. If we assume ε to be chosen as small as possible so that $D_\varepsilon T$ maps $\overline{K(2)}$ into itself, then $D_\varepsilon T$ will map some rank 1 matrix in $H(2)$ to a rank 1 matrix. It is easy to see that $D_\varepsilon T$ actually maps $K(2)$ into itself, and that $\varepsilon > 0$ (so that $D_\varepsilon T$ is invertible). Replacing T by $D_\varepsilon T$, and applying congruence and negation (if necessary), we henceforth assume that $T(E_{11}) = E_{11}$.

If the only rank 1 matrices mapped to rank 1 matrices by T are multiples of E_{11}, then an argument similar to that above indicates the existence of an ε, $0 < \varepsilon < 1$, such that $D_\varepsilon T$ preserves $K(2)$, $D_\varepsilon T(E_{11}) = E_{11}$, and $D_\varepsilon T(A)$ has rank 1, for some A which has rank 1 and is not a multiple of E_{11}. To within congruence, then, we may take T to fix both E_{11} and E_{22}.

Now for any $r \in \mathbf{R}$, $rE_{11} + (E_{12} + E_{21}) \in K(2)$. It follows that the (1, 2) and (2, 1) entries in $T(E_{12} + E_{21})$ are zero. The same is true, of course, for $T(iE_{12} - iE_{21})$. If the only rank 1 matrices mapped to rank 1 matrices by T are multiples of E_{11} or E_{22}, then we may argue once more that there is an ε, $0 < \varepsilon < 1$, such that $D_\varepsilon T$ preserves $K(2)$ and that there is a rank 1 matrix B such that $D_\varepsilon T(B)$ has rank 1, $D_\varepsilon T$ fixes both E_{11} and E_{22}, and E_{11}, E_{22}, and B are linearly independent. In addition, by performing a diagonal unitary congruence, we may also assume that B and $D_\varepsilon T(B)$ are real, and that T fixes $E_{12} + E_{21}$.

Now suppose that

$$T(iE_{12} - iE_{21}) = \begin{bmatrix} 0 & re^{i\theta} \\ re^{-i\theta} & 0 \end{bmatrix}. \tag{6.1}$$

Then for $u, v \in \mathbf{R}$,

$$T\begin{bmatrix} 0 & u+iv \\ u-iv & 0 \end{bmatrix} = \begin{bmatrix} 0 & u+rve^{i\theta} \\ u+rve^{-i\theta} & 0 \end{bmatrix}.$$

In order that T preserve $K(2)$, it is necessary that

$$|u + rve^{i\theta}| > |u + iv|$$

for any choice of $u, v \in \mathbf{R}$. This in turn, is equivalent to

$$\left\| \begin{bmatrix} 1 & r\cos\theta \\ 0 & r\sin\theta \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \right\| \geqq \left\| \begin{bmatrix} u \\ v \end{bmatrix} \right\| \tag{6.2}$$

for all $u, v \in \mathbf{R}$. For (6.2) to be satisfied, the minimum singular value of the matrix in (6.2) must be 1. This occurs if and only if $|r| \geqq 1$, and $\cos\theta = 0$, that is, $\theta = \pm\pi/2$. It then follows from (6.1) that $T = D_{1,r}$. This establishes Theorem 1.3. □

REFERENCES

[C] M.-D. CHOI, *Positive linear maps*, Proc. Sympos. Pure Math., 38 (1982), pp. 583–590.

[GN] B. C. GUNNING AND H. ROSSI, *Analytic Functions of Several Complex Variables*, Prentice-Hall, Englewood Cliffs, NJ, 1965.

[GLR1] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[GLR2] ———, *Matrices and Indefinite Scalar Products*, Birkhäuser, Boston, 1983.

[H] S. HELGASON, *Differential Geometry, Lie Groups and Symmetric Spaces*, Academic Press, New York, 1978.

[HJ] R. HORN AND C. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[HR] J. W. HELTON AND L. RODMAN, *Signature preserving maps on hermitian matrices*, Linear and Multilinear Algebra, 17 (1985), pp. 29–37.

[JP1] C. JOHNSON AND S. PIERCE, *Linear maps on hermitian matrices: the stabilizer of an inertia class*, Canad. Math. Bull., 28 (1985), pp. 401–404.

[JP2] ———, *Linear maps on hermitian matrices: the stabilizer of an inertia class*, II, Linear and Multilinear Algebra, 19 (1986), pp. 21–31.

[JR] C. JOHNSON AND L. RODMAN, *Convex sets of hermitian matrices with constant inertia*, SIAM J. Algebraic Discrete Methods, 6 (1985), pp. 351–359.

[L] S. LOJASIEWISZ, *Sur les ensembles semi-analytiques*, Actes, Congress. Internat. Math. Nice, 2 (1976), pp. 237–241.

[RR] A. C. M. RAN AND L. RODMAN, *Stability of invariant maximal semi-definite subspaces*, I, Linear Algebra Appl., 62 (1984), pp. 51–86.

[RU] A. C. M. RAN AND F. UHLIG, *A note on a new description of invariant maximal nonnegative subspaces in an indefinite inner product space*, Linear Algebra Appl., 71 (1985), pp. 273–274.

[T] R. C. THOMPSON, *The characteristic polynomial of a principal subpencil of a hermitian matrix pencil*, Linear Algebra Appl., 14 (1976), pp. 135–177.

[W] H. WHITNEY, *Elementary structure of real analytic varieties*, Ann. of Math., 66 (1957), pp. 545–556.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 4, October 1988

003

ON MINIMIZING THE SPECTRAL RADIUS OF A NONSYMMETRIC MATRIX FUNCTION: OPTIMALITY CONDITIONS AND DUALITY THEORY*

MICHAEL L. OVERTON† AND ROBERT S. WOMERSLEY‡

Abstract. Let $A(x)$ be a nonsymmetric real matrix affine function of a real parameter vector $x \in \mathscr{R}^m$, and let $\rho(x)$ be the spectral radius of $A(x)$. The article addresses the following question: Given $x_0 \in \mathscr{R}^m$, is $\rho(x)$ minimized locally at x_0, and, if not, is it possible to find a descent direction for $\rho(x)$ from x_0? If any of the eigenvalues of $A(x_0)$ that achieve the maximum modulus $\rho(x_0)$ are multiple, this question is not trivial to answer, since the eigenvalues are not differentiable at points where they coalesce. In the symmetric case, $A(x) = A(x)^T$ for all x, $\rho(x)$ is convex, and the question was resolved recently by Overton following work by Fletcher and using Rockafellar's theory of subgradients. In the nonsymmetric case $\rho(x)$ is neither convex nor Lipschitz, and neither the theory of subgradients nor Clarke's theory of generalized gradients is applicable. A new necessary and sufficient condition is given for $\rho(x)$ to have a first-order local minimum at x_0, assuming that all multiple eigenvalues of $A(x_0)$ that achieve the maximum modulus are nondefective. The optimality condition is computationally verifiable and involves computing "dual matrices." If the condition does not hold, the dual matrices provide information that leads to the generation of a descent direction. The result can be extended to the case where $\rho(x)$ is replaced by the maximum real part of the eigenvalues of $A(x)$. The authors use the eigenvalue perturbation theory of Rellich and Kato, which provides expressions for directional derivatives of $\rho(x)$. They also derive formulas for the codimension of manifolds on which certain eigenvalue structures of $A(x)$ are maintained; these are due to Von Neumann and Wigner and to Arnold. Finally, they discuss the much more difficult question of resolving optimality when $A(x_0)$ has a defective multiple eigenvalue achieving the maximum modulus $\rho(x_0)$.

Key words. nonsmooth optimization, nondifferentiable optimization, eigenvalue minimization, minimum spectral radius, nonconvex optimization

AMS(MOS) subject classifications. 65F99, 65K10, 90C25

1. Introduction.

Let $A(x)$ be a real $n \times n$ matrix affine function of $x = (\xi_1, \cdots, \xi_m)^T \in \mathscr{R}^m$, i.e.,

$$A(x) = A_0 + \sum_{k=1}^{m} \xi_k A_k, \tag{1.1}$$

where $\{A_k\}$ are given real $n \times n$ matrices. Define $\rho(x)$ to be the spectral radius of $A(x)$, i.e.,

$$\rho(x) = \max_{1 \leq i \leq n} |\lambda_i(x)|, \tag{1.2}$$

where $\lambda_i(x)$, $i = 1, \cdots, n$, are the (not necessarily distinct) eigenvalues of $A(x)$. Because $A(x)$ is real, the eigenvalues $\{\lambda_i(x)\}$ are either real or occur in complex conjugate pairs. In this paper we address the following question: Given $x_0 \in \mathscr{R}^m$, is $\rho(x)$ minimized locally by $x = x_0$, and if not, can we find a descent direction for ρ from x_0, that is, a direction $d \in \mathscr{R}^m$ such that $\rho(x_0 + \alpha d) < \rho(x_0)$ for sufficiently small $\alpha > 0$? There are several cases of increasing level of difficulty.

* Received by the editors August 20, 1987; accepted for publication (in revised form) February 1, 1988.

† Computer Science Department, Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. This author's research was supported in part by National Science Foundation grant DCR-85-02014, and took place while he was on sabbatical leave at the Centre for Mathematical Analysis and Mathematical Sciences Research Institute, Australian National University, Canberra ACT 2601, Australia.

‡ School of Mathematics, University of New South Wales, Kensington, New South Wales 2033, Australia.

If $A(x_0)$ has real eigenvalues of distinct modulus, $\rho(x)$ is differentiable, indeed analytic, at x_0 (see Kato (1984, p. 64)). The question is therefore answered by examining $\nabla\rho(x_0)$ and $\nabla^2\rho(x_0)$. The same is true when complex conjugate pairs of eigenvalues, each pair having different modulus, are permitted. For example, let $n = 2$, $m = 1$, and define

$$A(x) = \begin{bmatrix} 1+\xi_1 & 1 \\ 1 & 1-\xi_1 \end{bmatrix}.$$

Then $\lambda_{1,2}(x) = 1 \pm \sqrt{1+\xi_1^2}$, and $\rho(x)$ is minimized at $x_0 = [0]$, where $\nabla\rho = 0$ and $\nabla^2\rho$ is positive.

If several eigenvalues, not complex conjugates of each other, achieve the maximum modulus at x_0 but each eigenvalue is distinct, then $\rho(x)$ is simply the pointwise maximum function of several differentiable functions, and may be analyzed by standard min-max theory (see, e.g., Fletcher (1981, p. 175)).

Example 1.1. Let $n = 2$, $m = 1$, and define

$$A(x) = \begin{bmatrix} 1+\xi_1 & 1 \\ -\xi_1 & 1+\xi_1 \end{bmatrix}.$$

The eigenvalues are

$$\lambda_{1,2}(x) = 1 + \xi_1 \pm \sqrt{-\xi_1}$$

and the spectral radius is given by

$$\rho(x) = \begin{cases} -1-\xi_1+\sqrt{-\xi_1} & \text{if } \xi_1 \leqq -1, \\ 1+\xi_1+\sqrt{-\xi_1} & \text{if } -1 \leqq \xi_1 \leqq 0, \\ \sqrt{\xi_1^2+3\xi_1+1} & \text{if } \xi_1 \geqq 0 \end{cases}$$

(see Fig. 1.1). We see that at $x = [\xi_1] = -1$, the eigenvalues $\lambda_1(x)$ and $\lambda_2(x)$ have the same modulus, although they are distinct. The function $\rho(x)$ is a standard "max function" here; in particular, it is Lipschitz. On the other hand, at $x = 0$, the eigenvalues $\lambda_1(x)$ and $\lambda_2(x)$ coalesce and $\rho(x)$ has a completely different, non-Lipschitz, character. In fact, A is defective, i.e., not diagonalizable, at $x = 0$, and we say that $\lambda_1(x) = \lambda_2(x)$ is a defective eigenvalue. In general, even if $A(x)$ is nondefective at $x = x_0$, $\rho(x)$ is not differentiable at x_0 if $A(x_0)$ has multiple eigenvalues, and cannot be analyzed by standard min-max theory.

Besides showing the very different character of the two local minima, Fig. 1.1 also shows that, as typical with nonconvex problems, several local minima may occur and finding a global minimum would be very difficult in general. The example also shows that it is possible for $\rho(x)$ to have a smooth local maximum, so that the condition $\nabla\rho(x_0) = 0$ is not sufficient for f to have a local minimum at x_0.

An example with $m > 1$ gives additional insight.

Example 1.2. Let $n = 2$, $m = 2$, and define

$$A(x) = \begin{bmatrix} 2+\xi_1 & \xi_2 \\ 2\xi_1 & 2+\xi_1+\xi_2 \end{bmatrix}.$$

Figure 1.2 shows a contour plot of $\rho(x)$. There is no unconstrained local minimum of ρ. At the origin $x = [0, 0]^T$, $A(x)$ has a nondefective eigenvalue of multiplicity 2. Along the two lines $\xi_2 = 0$ and $\xi_2 = -8\xi_1$, except at the origin, $A(x)$ has a defective eigenvalue of multiplicity 2. These two lines divide the (ξ_1, ξ_2) plane into four quadrants; the ei-

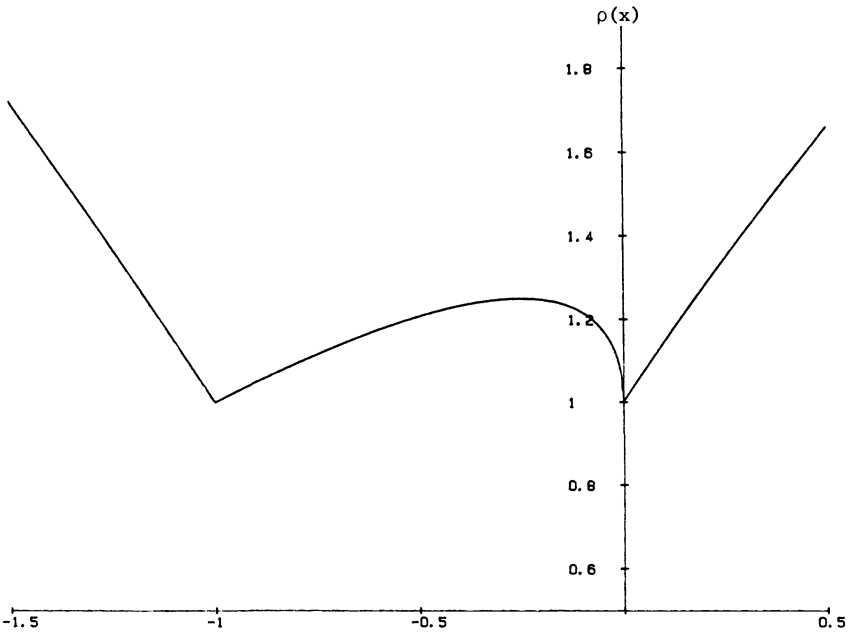


FIG. 1.1. *Plot of Example* 1.1.

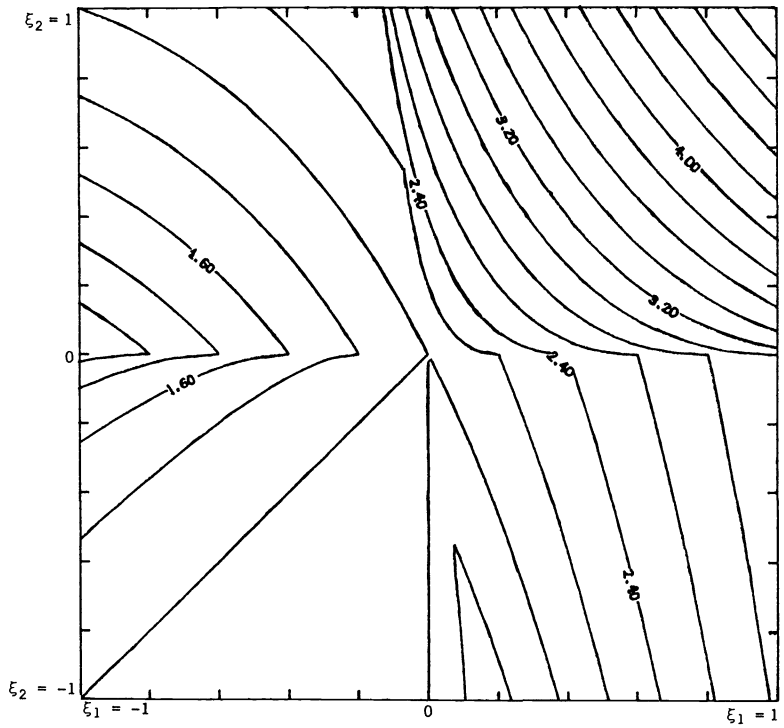


FIG. 1.2. *Contours of Example* 1.2.

genvalues are real and distinct in the top right and bottom left quadrants, and a complex conjugate pair in the other two quadrants. Note how the contours of ρ change sharply as they cross the defective manifold. This is because on the real side of the defective manifold, one of the eigenvalues is sharply increased by $O(\sqrt{\varepsilon})$ as the point x moves a distance ε away from the manifold, while on the complex side it is the imaginary part of the eigenvalue that is perturbed by $O(\sqrt{\varepsilon})$, which has only an $O(\varepsilon)$ effect on ρ. (The same effect is observed in Fig. 1.1 at $x = 0$.) Along lines passing through the origin, the function ρ is Lipschitz, but it is not Lipschitz along any other line in the (ξ_1, ξ_2) plane. Note that even if two vectors d_1 and d_2 are descent directions from the origin, a convex combination of d_1 and d_2 may be an ascent direction. We shall return to this phenomenon later.

Example 1.2 is not generic in the sense that a two-parameter family of matrices cannot be expected to have a nondefective multiple eigenvalue; this is explained in § 2. However, the example can be extended to three variables without changing its essential character by adding a term $\xi_3 A_3$ to $A(x)$. In that case the defective manifold becomes a cone instead of a pair of lines (see Arnold (1971, p. 40)). The eigenvalues of $A(x)$ are complex in the two disconnected "interior" parts of the cone and real elsewhere.

If $A(x) = A(x)^T$ for all x, $\rho(x)$ is a convex function and Rockafellar's theory of subgradients applies. In a recent paper, Overton (1988), following Fletcher (1985), has given verifiable optimality conditions for the symmetric case and shown how, if not optimal, a descent direction may always be obtained, even if this requires splitting a multiple eigenvalue. (There are exceptions in degenerate cases.) Both the optimality conditions and the method for obtaining descent directions involve an interesting duality theory. The same paper provides a practical, accurate algorithm for minimizing $\rho(x)$ in the symmetric case.

In the nonsymmetric case $\rho(x)$ is generally not convex and the problem is much more difficult. The main contribution of the present paper concerns the case where the (multiple) eigenvalues achieving the maximum modulus at x_0 are all nondefective. Even in this case, $\rho(x)$ is generally not Lipschitz at x_0, and hence the usual definition of the generalized gradient of Clarke (1975) is not applicable. However, the function ρ is Lipschitz at x_0 if its argument is restricted to the line $\{x_0 + \alpha d | \alpha \in \mathscr{R}\}$, for any $d \in \mathscr{R}^m$, and indeed the usual directional derivative of ρ (in the direction d) always exists. By considering this we are able to give a new necessary and sufficient condition for x_0 to be a local first-order minimizer of $\rho(x)$, excluding degenerate cases. The condition is computationally verifiable and involves computing "dual matrices." If the condition is found not to hold, the dual matrices are used to provide information that produces a descent direction, even if this requires splitting a multiple eigenvalue or making a multiple eigenvalue defective.

The paper is organized as follows. In the next section we derive formulas for the codimensions of manifolds defined by maintaining a given Jordan structure for $A(x)$. In the most general case, these formulas are due to Arnold (1971), (1983). In § 3 we characterize the directional derivative of $\rho(x)$. This derivation relies on the classic work of Kato and Rellich (see Kato (1984)). In § 4 we begin by summarizing the known optimality conditions for the symmetric case; we then derive new optimality conditions in the nonsymmetric case when only one multiple eigenvalue, which is nondefective at x_0, achieves the maximum modulus at x_0. In § 5 we extend this result to cover the case of several nondefective multiple eigenvalues achieving the same maximum modulus at x_0. In § 6 we briefly discuss the situation where a multiple eigenvalue achieving the maximum modulus is defective. The question of optimality seems very difficult to resolve in this case.

This paper is motivated by many applications. Perhaps the major source of applications is control engineering, where, for example, an optimal spectral radius value below

1 would represent system stability while a value greater than 1 would represent instability. See, for example, Mäkilä and Toivonen (1987) and Miller, Cochran, and Howze (1978) for applications where $A(x)$ is nonsymmetric; see Boyd (1988) and Kamenetskii and Pyatnitskii (1987) for applications where $A(x)$ is symmetric. Another source of applications is the design of iterative methods for solving linear systems of equations, where certain parameters must be chosen to minimize the spectral radius of the iteration matrix (see, for example, Young (1971)). The most well-known example is the SOR method, which depends on a single parameter ω whose optimal value is well known. More generally, we might consider a general preconditioner design problem. Since the latter application class generally involves nonlinear parameter dependence, the results of this paper cannot be applied directly. However, the results reported here will be an essential starting point for the analysis of problems where $A(x)$ is a nonlinear function. Other applications may involve constraints on the variables; it should be possible to extend the results given here to handling such constraints using standard Lagrange multiplier techniques.

As mentioned earlier, a practical algorithm is already available to minimize $\rho(x)$ in the symmetric case. We believe the results in this paper are an important first step towards the long-term goal of obtaining an efficient algorithm for the nonsymmetric case. There are many difficulties to be overcome before such a goal can be achieved. For example, even computing the Jordan form of $A(x)$ at a single point x is known to be a hard problem numerically, although there has been substantial progress in this direction in recent years (see Golub and Van Loan (1983) and Demmel (1983)).

It is important to note that the techniques used in this paper are also relevant to other functions of the eigenvalues $\lambda(x)$ besides the spectral radius. In fact, they could be used to analyze any real convex function of the eigenvalue function $\lambda(x) \in \mathscr{C}^m$. In our analysis of $\rho(x)$, we note that minimizing $\rho(x)$ is equivalent to minimizing

$$f(x) = \tfrac{1}{2}\rho(x)^2 = \tfrac{1}{2} \max_{1 \le i \le n} \lambda_i(x)\bar{\lambda}_i(x), \tag{1.3}$$

where $\bar{z}$ denotes the complex conjugate of $z \in \mathscr{C}$. Most of the analysis is then concerned with the nondifferentiable nature of $\lambda_i(x)$. Similarly, we can also consider minimizing another function that frequently arises in applications:

$$g(x) = \max_{1 \le i \le n} \operatorname{Re} \lambda_i(x) = \tfrac{1}{2} \max_{1 \le i \le n} (\lambda_i(x) + \bar{\lambda}_i(x)). \tag{1.4}$$

Of course, $\rho(x)$ and $g(x)$ are related to each other by exponential transformation of the matrix $A(x)$, but this is to be avoided numerically (Golub and Van Loan (1983)). In control engineering, for example, the form $g(x)$ arises when we consider stability of initial value problems; the form $\rho(x)$ arises when we consider discrete-time systems.

There is a large literature on extremal eigenvalue problems (see, for example, Nowasad (1968), Friedland (1978), and references therein). However, most of this work seems to be concerned with special problems that arise in infinite-dimensional spaces. The questions raised here do not seem to have been considered in detail previously.

2. The codimension of manifolds. An eigenvalue of multiplicity t is said to be nondefective (or semisimple) if the corresponding part of the Jordan form of the matrix is diagonal. Let x_0 be given, with $A(x_0)$ having a nondefective eigenvalue of multiplicity t, say $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$. What, generically, is the codimension of the manifold containing x_0 on which $A(x)$ has a nondefective multiple eigenvalue $\lambda_1(x) = \cdots = \lambda_t(x)$? This question was answered in the symmetric case by von Neumann and Wigner (1929) and, in the context of requiring a matrix to have a given rank, by Ledermann (1937), although the answer does not seem to be widely known. More recently, the

symmetric case was discussed by Friedland, Nocedal, and Overton (1987) in the context of inverse eigenvalue problems. Arnold (1971), (1983) answers the question in the general complex nonsymmetric case, including the defective case when nontrivial Jordan blocks must be considered. In this section we motivate and summarize these results, which are essential for a complete understanding of the later sections. We do not give a rigorous derivation, for which the reader is referred to Arnold's work.

First assume that $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$ is real and that the other eigenvalues of $A(x_0)$ are real and distinct. For x to lie in the desired manifold, we require

$$A(x)Q = Q\Lambda, \qquad \Lambda = \begin{bmatrix} \lambda I_t & \\ & \Lambda_2 \end{bmatrix}, \tag{2.1}$$

where Q is a nonsingular real matrix, I_t is the identity matrix of order t, and Λ_2 is a real diagonal matrix of order $n - t$. (None of the eigenvalues can become complex near enough to x_0 since the only multiple eigenvalue is being preserved.) We may view (2.1) as $h_1 = n^2$ equations that restrict x; but we have introduced additional variables Q and Λ. These variables are correctly counted as follows. There are $h_2 = n - t + 1$ variables in Λ. The matrix Q has $h_3 = n^2$ components, but not all n^2 degrees of freedom are useful in satisfying (2.1). Let $Q = [Q_1, Q_2]$, where the columns of Q_1 correspond to $\lambda_1(x) = \cdots = \lambda_t(x)$. We may postmultiply Q_1 by any nonsingular $t \times t$ matrix, and postmultiply Q_2 by any nonsingular diagonal matrix, without affecting (2.1). Let $h_4 = t^2$ and $h_5 = n - t$; therefore, we see that the total number of introduced variables useful in solving (2.1) is

$$h_2 + h_3 - h_4 - h_5 = n^2 - t^2 + 1.$$

The codimension of the desired manifold is obtained by subtracting this from h_1, the number of equations in (2.1), giving

$$c_N(t) = h_1 - h_2 - h_3 + h_4 + h_5 = t^2 - 1. \tag{2.2}$$

Since this manifold is embedded in $\mathscr{R}^m$, and the codimension describes the number of degrees of freedom restricted by requiring x to be in the manifold, the dimension of the manifold is $m + 1 - t^2$. For example, if $t = 2$ and $m = 3$, the dimension of the manifold is zero, i.e., a three-parameter matrix family $A(x)$ generically has only a single point x_0, where $A(x_0)$ has a nondefective multiple eigenvalue. Of course, this argument is generic and there are exceptions in degenerate cases.

A similar argument for the symmetric case ($A(x) = A(x)^T$ for all x) gives the Von Neumann–Wigner result $h_1 = n(n + 1)/2$, $h_2 = n - t + 1$, $h_3 = n(n - 1)/2$ (since Q is orthogonal), $h_4 = t(t - 1)/2$, $h_5 = 0$ (since Q is already restricted to being orthogonal by h_3), so

$$c_S(t) = \frac{t(t+1)}{2} - 1. \tag{2.3}$$

Von Neumann and Wigner also derived the codimension for the case that $A(x)$ is complex but Hermitian for all x, where we continue to view $A(x)$ as a function of *real* variables; thus (2.1) is n^2 real equations, namely $n(n - 1)/2$ complex off-diagonal equations and n real diagonal equations. We then obtain the same formula as (2.2).

Returning to the real nonsymmetric case, if $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$ is real but we allow the other eigenvalues to be complex, the codimension (2.2) does not change. This is because Λ_2 and Q_2, although complex, consist of complex conjugate pairs.

If the multiple eigenvalue $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$ is one of a complex conjugate pair, we require

$$A(x)Q = Q\Lambda, \qquad \Lambda = \begin{bmatrix} \lambda_1 I_t & & \\ & \bar{\lambda}_1 I_t & \\ & & \Lambda_2 \end{bmatrix},$$

where $Q = [Q_1, \bar{Q}_1, Q_2]$, and Λ_2 is a diagonal matrix of order $n - 2t$. We obtain $h_1 = n^2$, $h_2 = n - 2t + 2$, $h_3 = n^2$, $h_4 = 2t^2$, and $h_5 = n - 2t$, i.e.,

$$c_C(t) = 2t^2 - 2. \tag{2.4}$$

Thus the codimension is the same as if two real multiple eigenvalues, each of multiplicity t, were to be preserved separately.

Suppose we require $r + s$ nondefective multiple eigenvalues to have the same modulus, where r of them are real with respective multiplicities, $t_1, \cdots, t_r$, and s of them are complex with positive imaginary part and respective multiplicity $t_{r+1}, \cdots, t_{r+s}$. (Note $r \leqq 2$.) Then the codimension of the manifold along which multiplicities are preserved and all eigenvalues have the same modulus is

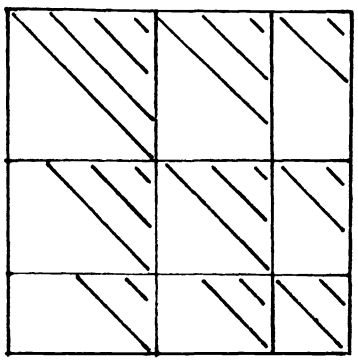
$$\begin{aligned} c_G(t_1, \cdots, t_r; t_{r+1}, \cdots, t_s) &= \sum_{j=1}^{r} (t_j^2 - 1) + \sum_{j=r+1}^{r+s} (2t_j^2 - 2) + (r + s - 1) \\ &= \sum_{j=1}^{r} t_j^2 + 2\sum_{j=r+1}^{r+s} t_j^2 - s - 1, \end{aligned} \tag{2.5}$$

reflecting the fact that $(r + s - 1)$ additional restrictions are being placed on the moduli.

Now let us drop the assumption that $\lambda_1(x_0)$ is nondefective. Assume $A(x_0)$ has a real multiple eigenvalue $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$, corresponding to Jordan blocks of size $u_1 \geqq u_2 \geqq \cdots \geqq u_p$, $1 \leqq p \leqq t$. We are interested in the dimension of the manifold passing through x_0 along which the same Jordan structure is maintained. For x to lie in the manifold, we require that

$$A(x)Q = QJ, \qquad J = \begin{bmatrix} J_1 & \\ & \Lambda_2 \end{bmatrix},$$

where $Q = [Q_1, Q_2]$ is any nonsingular matrix, Λ_2 is diagonal of order $n - t$, and J_1 is the desired Jordan form. We have $h_1 = n^2$, $h_2 = n - t + 1$, $h_3 = n^2$, and $h_5 = n - t$ as before. To determine h_4 we need to answer the following question: What class of matrices commute with J_1? If J_1 equals $\lambda_1 I$, the answer is all $t \times t$ matrices; if J_1 is a single Jordan block, the answer is all $t \times t$ upper triangular Toeplitz matrices. In general, the answer is given by Arnold (1971, p. 34), namely matrices of the following form:



Here the block partitioning conforms to the Jordan block partitioning of J_1, and each block is an upper triangular (rectangular) Toeplitz matrix. The example shown here corresponds to $u_1 = 4$, $u_2 = 3$, $u_3 = 2$. The number of variables in such a matrix is

$$h_4 = u_1 + 3u_2 + 5u_3 + \cdots + (2p+1)u_p.$$

We therefore obtain the codimension

$$c_D(u_1, \cdots, u_p) = u_1 + 3u_2 + 5u_3 + \cdots + (2p+1)u_p - 1. \tag{2.6}$$

Note that, as before, the codimension is independent of n. We have

$$c_D(1, 1, \cdots, 1) = t^2 - 1 = c_N(t)$$

and the codimension for a single Jordan block is

$$c_D(t) = t - 1. \tag{2.7}$$

The arguments given here are not rigorous; in particular we have not attempted to prove independence of the various restricting equations. For a full derivation, see Arnold (1971).

3. Directional derivatives. Let x_0 be given with $A(x_0)$ having a nondefective multiple eigenvalue $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$. In general the eigenvalues $\lambda_i(x)$, $i = 1, \cdots, t$, are not Lipschitz functions even at x_0. For example, let

$$A(x) = \begin{bmatrix} 1 & \xi_1 \\ \xi_2 & 1 \end{bmatrix}$$

so that

$$\lambda_{1,2}(x) = 1 \pm \sqrt{\xi_1 \xi_2}.$$

Given any ball of radius $\varepsilon > 0$ around $x_0 = [0, 0]^T$, let $x_1 = [\varepsilon/\sqrt{2}, 0]^T$ and $x_2 = [\varepsilon/\sqrt{2}, \delta]^T$, where $\delta > 0$. Both x_1 and x_2 lie in the given ball if $\delta \leqq \varepsilon/\sqrt{2}$, but

$$|\lambda_i(x_1) - \lambda_i(x_2)|$$

cannot be bounded by $K\delta$ for any constant K independent of δ. This contradicts the definition of a Lipschitz function (which may be found in, e.g., Clarke (1983)). Of course, the eigenvalues $\lambda_i(x)$ are always continuous functions, regardless of x_0, provided a consistent ordering is used.

Although the eigenvalues $\lambda_i(x)$, $i = 1, \cdots, t$, are not Lipschitz with respect to several variables, they may be ordered so that they are locally continuously differentiable along any line passing through x_0. This follows from the classical eigenvalue perturbation theory of Rellich and Kato. Before stating the result let us introduce some notation. Let Q_1 be an $n \times t$ matrix whose columns are independent right eigenvectors of $A(x_0)$ corresponding to $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$ and let P_1^T be a $t \times n$ matrix whose rows are corresponding independent left eigenvectors. We may normalize P_1 so that

$$P_1^T Q_1 = I_t \tag{3.1}$$

and we then have

$$P_1^T A(x_0) Q_1 = \lambda_1(x_0) I_t. \tag{3.2}$$

The quantity $Q_1 P_1^T$ is called the eigenprojection for $\lambda_1(x_0)$. Define the $t \times t$ matrices

$$B_k = P_1^T A_k Q_1, \qquad k = 1, \cdots, m, \tag{3.3}$$

where A_k is given by (1.1). Note that if $\lambda_1(x_0)$ is real, all of P_1, Q_1, and B_k, $k = 1, \cdots, m$, are also real, but if $\lambda_1(x_0)$ is complex, all these matrices generally will also be complex. If $\lambda_1(x_0)$ is complex, it has an associated complex conjugate multiple eigenvalue, with corresponding eigenvector matrices $\bar{P}_1$, $\bar{Q}_1$, and

$$\bar{B}_k = \bar{P}_1^T A_k \bar{Q}_1, \qquad k = 1, \cdots, m$$

corresponding to (3.3).

Now define the directional derivative of $\lambda_i(x)$ in the direction $d = [\delta_1, \cdots, \delta_m]^T \in \mathscr{R}^m$ by

$$\lambda_i'(x_0; d) = \lim_{\alpha \to 0^+} \frac{\lambda_i(x_0 + \alpha d) - \lambda_i(x_0)}{\alpha}. \tag{3.4}$$

LEMMA 3.1. *We have*

$$\lambda_i'(x_0; d) = \mu_i, \qquad i = 1, \cdots, t, \tag{3.5}$$

where $\{\mu_i\}$ *are the eigenvalues of*

$$B(d) = \sum_{k=1}^{m} \delta_k B_k. \tag{3.6}$$

Proof. See Kato (1984, p. 81) and preceding pages for the proof. Note that, although we assume that $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$ is nondefective, we do not assume that $B(d)$ is nondefective.

Remark. It is useful to motivate the result as follows. Suppose for simplicity that $B(d)$ is nondefective, and let its eigensystem be

$$B(d) = Z D Y^T, \tag{3.7}$$

where Y, Z are nonsingular $t \times t$ matrices, $Y^T Z = I_t$, and D is diagonal with entries $\{\mu_i\}$. We have

$$Y^T P_1^T A(x_0 + \alpha d) Q_1 Z = \lambda_1(x_0) + \alpha D. \tag{3.8}$$

If $t = n$, this proves the lemma, since (3.8) gives the eigensystem of $A(x_0 + \alpha d)$, with linear eigenvalues $\lambda_1(x_0) + \alpha\mu_i$. On the other hand, if $t = 1$, the lemma is trivial since μ_1 is the inner product of d with the gradient of the differentiable function $\lambda_1(x)$, namely $[p_1^T A_1 q_1, \cdots, p_1^T A_m q_1]^T$. More generally, suppose that $1 < t < n$. Then (3.8) represents a generalized Rayleigh quotient, the key point being that the right-hand side is diagonal. Thus the diagonal entries approximate the first t eigenvalues of $A(x_0 + \alpha d)$, and the columns of $Q_1 Z$ (respectively, the rows of $Y^T P_1^T$) are the particular right (respectively, left) eigenvectors of $A(x_0)$ to which the right (left) eigenvectors of $A(x_0 + \alpha d)$ generally converge as $\alpha \to 0$. (If the $\{\mu_i\}$ are not distinct, the corresponding eigenvectors need not converge.)

Now let us turn to the functions $f(x)$ and $g(x)$ defined by (1.3) and (1.4); it is easier to work with $f(x) = \frac{1}{2}\rho(x)^2$ than directly with $\rho(x)$. Note that as long as $f'(x_0; d)$ exists with $f(x_0) \neq 0$, the quantity $\rho'(x_0; d)$ exists and is related by

$$\rho'(x_0; d) = \frac{f'(x_0; d)}{\rho(x_0)}.$$

LEMMA 3.2. *Suppose that* $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$ *is a real nondefective eigenvalue, and that all other eigenvalues of* $A(x_0)$ *have smaller modulus than* $|\lambda_1(x_0)|$. *Then for any* $d \in \mathcal{R}^m$

$$f'(x_0; d) = \lambda_1(x_0) \max_{1 \leq i \leq t} \operatorname{Re} \mu_i,$$

where, as before, $\{\mu_i\}$ *are the eigenvalues of* $B(d)$.

Proof. It is clear that

$$f'(x_0; d) = \max_{1 \leq i \leq t} f_i'(x_0; d),$$

where

$$f_i(x_0; d) = \tfrac{1}{2}\lambda_i(x)\bar{\lambda}_i(x).$$

Now

$$f_i'(x_0; d) = \tfrac{1}{2}(\lambda_i(x_0)\bar{\lambda}_i'(x_0; d) + \bar{\lambda}_i(x_0)\lambda_i'(x_0; d)),$$

so the result follows from Lemma 3.1, since $\lambda_1(x_0)$ is real. □

LEMMA 3.3. *Suppose that* $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$ *is a nondefective eigenvalue, and that all other eigenvalues of* $A(x_0)$ *have smaller real part. Then*

$$g'(x_0; d) = \max_{1 \leq i \leq t} \operatorname{Re} \mu_i,$$

where again $\{\mu_i\}$ *are the eigenvalues of* $B(d)$.

Proof. The proof is straightforward.

More generally, consider the case where several different eigenvalues achieve the maximum modulus or the maximum real part, respectively. It is convenient to change notation as follows. Let $\lambda_{jl}(x)$ denote the eigenvalues of $A(x)$ with the following properties:

(i) $\lambda_{j1}(x_0) = \cdots = \lambda_{jt_j}(x_0)$, for $j = 1, \cdots, r$, is a real nondefective multiple eigenvalue of $A(x_0)$ with multiplicity t_j.

(ii) $\lambda_{j1}(x_0) = \cdots = \lambda_{jt_j}(x_0)$, for $j = r+1, \cdots, r+s$, is a complex nondefective multiple eigenvalue of $A(x_0)$ with multiplicity t_j and positive imaginary part.

(iii) $\{\lambda_{j1}(x_0)\}$, $j = 1, \cdots, r+s$, are distinct quantities with, in the case of minimizing the spectral radius, the same modulus $\rho(x_0) = \sqrt{2f(x_0)}$, or, in the case of minimizing the maximum real part, the same real part $g(x_0)$. These eigenvalues are said to be *active*. The complex conjugates $\bar{\lambda}_{j1}(x_0)$, $j = r+1, \cdots, r+s$, are also active, so there are a total of $r + 2s$ distinct active eigenvalues. All other eigenvalues of $A(x_0)$ are inactive, i.e., they have smaller modulus or smaller real part, respectively. (Note that $r \leqq 2$.)

Now, for $j = 1, \cdots, r+s$, define Q_j, P_j^T as matrices whose columns (respectively, rows) are independent right (respectively, left) eigenvectors of $A(x_0)$ corresponding to $\lambda_{j1}(x_0) = \cdots = \lambda_{jt_j}(x_0)$, with $P_j^T Q_j = I_{t_j}$. Define the $t_j \times t_j$ matrix

$$B_k^{(j)} = P_j^T A_k Q_j, \quad k = 1, \cdots, m, \quad j = 1, \cdots, r+s. \tag{3.9}$$

LEMMA 3.4. *Let* $A(x_0)$ *have nondefective active eigenvalues with respect to the function* $f(x)$. *For any* $d = [\delta_1, \cdots, \delta_m]^T \in \mathcal{R}^m$,

$$f'(x_0; d) = \max_{1 \leq j \leq r+s} \max_{1 \leq l \leq t_j} \operatorname{Re}(\bar{\lambda}_{j1}(x_0)\mu_{jl}),$$

where μ_{jl}, $l = 1, \cdots, t_j$ *are the eigenvalues of*

$$B^{(j)}(d) = \sum_{k=1}^{m} \delta_k B_k^{(j)}. \tag{3.10}$$

Proof. It is clear that

$$f'(x_0;d) = \max_{1 \le j \le r+s} \max_{1 \le l \le t_j} f'_{jl}(x_0;d),$$

where

$$f_{jl}(x) = \tfrac{1}{2}\lambda_{jl}(x)\bar{\lambda}_{jl}(x).$$

Since

$$f'_{jl}(x_0;d) = \tfrac{1}{2}(\lambda_{jl}(x_0)\bar{\lambda}'_{jl}(x_0;d) + \bar{\lambda}_{jl}(x_0)\lambda'_{jl}(x_0;d)) \tag{3.11}$$

the result follows from Lemma 3.1. □

LEMMA 3.5. *Let $A(x_0)$ have nondefective active eigenvalues with respect to the function $g(x)$. For any $d \in \mathcal{R}^m$,*

$$g'(x_0;d) = \max_{1 \le j \le r+s} \max_{1 \le l \le t_j} \operatorname{Re} \mu_{jl},$$

where μ_{jl}, $l = 1, \cdots, t_j$, are the eigenvalues of $B^{(j)}(d)$ defined by (3.10).

Proof. The proof is straightforward.

We complete this section with the definition of a matrix inner product that will be needed in § 4. Following Fletcher (1985), define

$$A:B = \operatorname{tr} A^T B \tag{3.12}$$

for any real rectangular matrices A and B with the same dimension.

LEMMA 3.6. $XAY^T:B = A:X^TBY$.

Proof. The proof is straightforward.

4. Optimality conditions in the case of one active nondefective multiple eigenvalue. Assume that $A(x_0)$ has one active multiple eigenvalue that is real, nonzero, and nondefective, and that we denote by $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$, reverting to our original notation. Let us define $d \in \mathcal{R}^m$ to be a descent direction for f from x_0 if $f'(x_0; d) < 0$. If no such direction exists, f is said to have a *first-order local minimum* at x_0. We wish to give a procedure for determining whether f has a first-order local minimum at x_0 and, if it does not, for obtaining a descent direction.

It is useful to first consider the symmetric case.

(1) Symmetric case ($A(x) = A(x)^T$ for all x).

In this case the eigenvalues $\lambda_i(x)$ are always real, the eigenvector matrix Q is orthogonal, $P_1 = Q_1$, and $B_k = Q_1^T A_k Q_1$. Furthermore, $f(x)$ and $\rho(x)$ are convex; this follows from Fletcher (1985, p. 510).

THEOREM 4.1. *Define the set*

$$\Omega = \{v = [\nu_1, \cdots, \nu_m]^T \in \mathcal{R}^m \mid \text{there exists a symmetric positive semidefinite } t \times t \text{ matrix } U \text{ satisfying } \operatorname{tr} U = 1, \lambda_1(x_0)U:B_k = \nu_k, k = 1, \cdots, m\}.$$

(The matrix inner product operator ":" was defined by (3.12).) A necessary and sufficient condition for x_0 to minimize f is that $0 \in \Omega$.

Proof. Let $v \in \Omega$, let $d = [\delta_1, \cdots, \delta_m]^T \in \mathcal{R}^m$, and let the eigenvalue decomposition of the $t \times t$ symmetric matrix $B(d) = \sum_{k=1}^m \delta_k B_k$ be given by $B = ZMZ^T$, where Z is orthogonal and $M = \operatorname{diag}(\mu_i)$. We have

$$v^T d = \lambda_1(x_0) \sum_{k=1}^m \delta_k U:B_k$$

$$= \lambda_1(x_0)\, U:ZMZ^T.$$

Therefore

$$\sup_{v \in \Omega} v^T d = \lambda_1(x_0) \sup_U U : Z M Z^T, \tag{4.1}$$

where the second "sup" is taken over all $t \times t$ symmetric positive semidefinite matrices U with tr $U = 1$. Since Z is orthogonal and U is symmetric, without loss of generality we may write (4.1) as

$$\lambda_1(x_0) \sup_U U : M = \lambda_1(x_0) \sup_U \sum_{i=1}^{m} U_{ii} \mu_i \tag{4.2}$$

(see Lemma 3.6). Now U cannot have negative diagonal elements, and it has trace equal to one, so we see from (4.1), (4.2) that

$$\begin{aligned} \sup_{v \in \Omega} v^T d &= \lambda_1(x_0) \max_{1 \leq i \leq t} \mu_i \\ &= f'(x_0; d) \end{aligned} \tag{4.3}$$

by Lemma 3.2. It follows that if $0 \in \Omega$,

$$f'(x_0; d) \geqq 0 \quad \forall d \in \mathscr{R}^m,$$

i.e., x_0 minimizes f. On the other hand, if $0 \notin \Omega$, then by the separating hyperplane theorem and the convexity of Ω, there exists d with $v^T d < 0$ for all $v \in \Omega$, i.e., d is a descent direction by (4.3). (For a statement of the separating hyperplane theorem, see Rockafellar (1970, p. 95).)

Remark. This theorem was proved in Overton (1988). The proof here is more direct, since it does not use Rockafellar's theory of subgradients, but only the separating hyperplane theorem. Nonetheless, the proof technique is similar to those used in the theory of subgradients, and it is doubtful whether the theorem would have been obtained without the motivation of that theory (and also the paper of Fletcher (1985)).

COROLLARY. $\Omega = \partial f(x_0)$, *the subdifferential of the convex function f as defined by Rockafellar* (1970).

Proof. The proof follows from (4.3).

Remark. Because f is convex, there is no distinction between "first-order local minimum" and "minimum."

Remark. The matrix U is called the dual matrix (or Lagrange matrix), and it plays the role of Lagrange multipliers familiar from constrained optimization.

We now discuss the generation of descent directions if x_0 is not optimal. There are three cases.

(1A) Symmetric case, assuming $I_t \in \text{Span}\{B_1, \cdots, B_m\}$.

In this case we simply solve

$$\lambda_1(x_0) \sum_{k=1}^{m} \delta_k B_k = -I_t. \tag{4.4}$$

By Lemma 3.2, $d = [\delta_1, \cdots, \delta_m]^T$ is a descent direction for f. Furthermore, all the eigenvalues $\lambda_1(x), \cdots, \lambda_t(x)$ decrease at the same rate along d; that is, the eigenvalue does not split to first order. This case holds generically if $m \geqq t(t+1)/2$, i.e., $m > c_S(t)$, i.e., the generic dimension of the manifold defined by

$$\lambda_1(x) = \cdots = \lambda_t(x) \tag{4.5}$$

is greater than zero.

(1B) Symmetric case, assuming (1A) does not apply and the set $\{I_t, B_1, \cdots, B_m\}$ has full rank $t(t+1)/2$.

This case holds generically when $m = c_S(t) = t(t+1)/2 - 1$, i.e., the manifold defined by (4.5) is the single point x_0. It also holds if $m > c_S(t)$, but f is minimized on the manifold (4.5) at x_0. To make further progress we must split the multiple eigenvalue λ_1. Solve for the dual matrix $U = U^T$ in the linear system

$$\text{tr } U = 1, \quad \lambda_1(x_0) U : B_k = 0, \quad k = 1, \cdots, m. \tag{4.6}$$

This is a system of $m+1$ equations in $t(t+1)/2$ unknowns. Although it is possible that the $\{B_k\}$ are not independent, (4.6) has a unique solution U in view of the homogeneity of all equations except the trace equation (which is equivalent to $I_t : U = 1$). If $0 \notin \Omega$, i.e., x_0 is not optimal, it follows that U is not positive semidefinite.

THEOREM 4.2. *Assume* $0 \notin \Omega$, *so that* U *has an eigenvalue* $\theta < 0$. *Let* $z \in \mathcal{R}^t$ *be a corresponding normalized eigenvector of* U. *Solve for* $[\delta_0, \delta_1, \cdots, \delta_m]^T \in \mathcal{R}^{m+1}$ *in*

$$\delta_0 I_t + \lambda_1(x_0) \sum_{k=1}^{m} \delta_k B_k = -zz^T. \tag{4.7}$$

Then $d = [\delta_1, \cdots, \delta_m]^T$ *is a descent direction.*

Proof. The linear system (4.7) is solvable by assumption, although if $\{B_k\}$ are not independent, d is not unique. Taking an inner product of U with (4.7) we obtain

$$\delta_0 \text{ tr } U + \lambda_1(x_0) \sum_{k=1}^{m} \delta_k U : B_k = -U : zz^T,$$

i.e.,

$$\delta_0 = -\theta > 0$$

by (4.6). From (4.7) and Lemma 3.2, $f'(x_0; d)$ is the maximum eigenvalue of the symmetric matrix $-zz^T - \delta_0 I_t$. The eigenvalues of this matrix are $(-1 + \theta, \theta, \cdots, \theta)$, so $f'(x_0; d) < 0$.

Remark. This theorem was given by Overton (1988). The proof here is slightly different. The theorem shows that we can progress by splitting the multiple eigenvalue while maintaining multiplicity $t - 1$ (to first order). This is analogous to moving off a single active constraint in the context of constrained optimization. Note that it is the dual matrix U that provides information leading to a descent direction, just as negative Lagrange multipliers provide similar information in constrained optimization. Note in particular that the coefficient matrix of the left-hand side of the linear system (4.6), which defines the dual matrix, is the transpose of the coefficient matrix of the linear system (4.7), which gives the descent direction.

(1C) Symmetric case, where neither (1A) nor (1B) applies.

Although this applies generally if $m < c_S(t)$, such cases are degenerate in the sense that, generically, a point x_0 satisfying (4.5) will not exist. In such degenerate situations, verifying optimality or finding a descent direction is very difficult, just as it is in the much simpler case of linear programming. We may be able to solve (4.6), but the dual matrix U is not uniquely defined and generally (4.7) will not be solvable. Theorem 4.1 still applies, so x_0 is optimal if and only if there exists a dual matrix U with the required properties. However, because the solution to (4.6) is not unique, finding such a matrix U may be very difficult.

We now turn to the nonsymmetric problem. We first dispose of the trivial case.

(2A) Nonsymmetric case, assuming $I_t \in \text{Span}\{B_1, \cdots, B_m\}$.

A descent direction is obtained by solving (4.4). The eigenvalue is not split (to first order). This case holds generically if $m > c_N(t) = t^2 - 1$.

(2B) Nonsymmetric case, assuming (2A) does not hold and the set $\{I_t, B_1, \cdots, B_m\}$ has full rank t^2.

This case holds generically when $m = c_N(t)$, i.e., the manifold defined by maintaining the nondefective multiple eigenvalue is the single point x_0. It also holds if $m > c_N(t)$, but f is minimized on the manifold at x_0. To make further progress we must either split the multiple eigenvalue λ_1 or make it defective.

Our initial work on this problem involved the following set, intended to generalize the subdifferential Ω to the nonconvex case. Define the (nonconvex) set Ψ by

$$\Psi = \{v = [v_1, \cdots, v_m]^T \in \mathscr{R}^m \mid \text{there exists a real } t \times t \text{ diagonalizable matrix } U \text{ with real nonnegative eigenvalues satisfying tr } U = 1, \lambda_1(x_0)U : B_k = v_k, k = 1, \cdots, m\}.$$

However, it is not the case that (4.3) holds when we substitute Ψ for Ω on the left-hand side. On the contrary,

$$\sup_{v \in \Psi} v^T d = \infty.$$

The point where the proof of Theorem 4.1 breaks down in the nonsymmetric case is that U can have negative diagonal elements, even though it is similar to a nonnegative diagonal matrix with trace equal to 1.

Nonetheless, it is true that $0 \in \Psi$ is a necessary condition for x_0 to minimize f. A weaker result which is easier to show, following the lines of Theorem 4.1, is that $0 \in \text{Conv } \Psi$ is a necessary condition for optimality, but this is of no interest since it turns out that $\text{Conv } \Psi = \mathscr{R}^m$. We note that if we were to apply the usual definition of Clarke's generalized gradient (Clarke (1983, p. 10)) to f, ignoring the fact that f is not Lipschitz, we would obtain $\partial f(x_0) = \Re^m$. Rockafellar has extended the definition of the generalized gradient to the non-Lipschitz case, but this apparently still gives $\partial f(x_0) = \Re^m$ for our function f (Rockafellar (1985), Burke (1987)).

It may be worth noting at this point that there cannot exist any set $\tilde{\Psi}$, convex or not, such that

$$\sup_{v \in \tilde{\Psi}} v^T d = f'(x_0; d) \quad \text{for all } d \in \mathscr{R}^m.$$

The existence of such a set would contradict the possibility of the existence of descent directions whose convex combination is an ascent direction, which was noted in Example 1.2.

To show that $0 \in \Psi$ is a necessary condition for x_0 to minimize f, first observe that, as in case (1B), the linear system (4.6) is solvable, although since the matrices are nonsymmetric, it is now a system of $m + 1$ equations in t^2 unknowns, namely the elements of the dual matrix U. If U has a negative real eigenvalue, we can obtain a descent direction by solving (4.7), replacing the right-hand side by yz^T, where z and y^T are, respectively, right and left eigenvectors for the negative eigenvalue of U. If U has complex eigenvalues or is defective, we can also find a descent direction by appropriate choice of the right-hand side of (4.7). In view of the subsequent remarks, there is no need to elaborate on this further.

We now show that the set Ψ is too large to be useful and that a necessary and sufficient optimality condition can be obtained from using a smaller set. Define

$$\Phi = \{ v = [v_1, \cdots, v_m]^T \in \mathscr{R}^m \mid U = \frac{1}{t} I_t, \lambda_1(x_0) U : B_k = v_k, k = 1, \cdots, m \},$$

i.e., Φ consists of the single point $v = (\lambda_1(x_0)/t)[\operatorname{tr} B_1, \cdots, \operatorname{tr} B_m]^T$.

THEOREM 4.3. *A necessary and sufficient condition for f to have a first-order local minimum at x_0 is that* $0 \in \Phi$.

Remark. The theorem does not require the assumption that $\lambda_1(x_0) \neq 0$. However, it is convenient to assume throughout that $\lambda_1(x_0) \neq 0$, as stated at the beginning of the section, so that (4.6) remains solvable. With this assumption, $0 \in \Phi \Leftrightarrow \operatorname{tr} B_k = 0$, $k = 1, \cdots, m$.

Proof. Define U by solving (4.6). The theorem states that $U = (1/t)I_t$ if and only if f has a first-order local minimum at x_0. First suppose that $U = (1/t)I_t$, and suppose also that x_0 is not a first-order local minimizer, i.e., there exists a descent direction $d \in \mathscr{R}^m$. By Lemma 3.2, this implies that

$$\lambda_1(x_0) \operatorname{Re} \mu_i < 0, \qquad i = 1, \cdots, t,$$

where μ_i are the eigenvalues of $B(d)$. Because $\lambda_1(x_0)$ is real, $B(d)$ is also real, so this implies $\operatorname{tr} \lambda_1(x_0) B(d) < 0$. However, this is a contradiction, since $U = (1/t)I_t$ implies $\lambda_1(x_0) \operatorname{tr} B_k = 0$, $k = 1, \cdots, m$.

Now suppose that f has a first-order local minimum at x_0, but that $U \neq (1/t)I_t$. The latter assumption implies that there exists a $t \times t$ real matrix E with zero eigenvalues such that $U : E \neq 0$, namely, one of the following $t^2 - 1$ linearly independent defective matrices:

$$e_p e_q^T, \quad p, q = 1, \cdots, t, \quad p \neq q$$

or

$$e_p e_p^T - e_{p+1} e_{p+1}^T + e_p e_{p+1}^T - e_{p+1} e_p^T, \qquad p = 1, \cdots, t-1.$$

Here e_p denotes the pth column of I_t. Now solve the following linear system for $[\delta_0, \delta_1, \cdots, \delta_m]^T = [\delta_0, d^T]^T \in \mathscr{R}^{m+1}$:

$$\delta_0 I_t + \lambda_1(x_0) \sum_{k=1}^{m} \delta_k B_k = E. \tag{4.8}$$

(This system is a nonsymmetric version of (4.7), and therefore the coefficient matrix of the left-hand side is the transpose of that in the nonsymmetric version of (4.6).) Taking an inner product of U with (4.8), we get

$$\delta_0 = U : E \neq 0. \tag{4.9}$$

But by (4.8) and Lemma 3.2, $f'(x_0; d)$ is the largest real part of the eigenvalues of $E - \delta_0 I_t$, i.e., $-\delta_0$. This contradicts the assumption that a descent direction does not exist, since if $\delta_0 < 0$ we may replace $[\delta_0, d^T]^T$ by $-[\delta_0, d^T]^T$. □

Any direction d that preserves the multiple eigenvalue $\lambda_1 = \cdots = \lambda_t$ (to first order) by making it defective (to first order) has the property that $f'(x_0; -d) = -f'(x_0; d)$, since all the active eigenvalues have the same first-order charge. It follows that either d or $-d$ is a descent direction unless the first-order charge is zero; Theorem 4.3 states that this happens for all such "defective" directions if and only if $U = (1/t)I_t$. An example of this is the following.

Example 4.1. Let $n = 2$, $m = 3$, and define

$$A(x) = \begin{bmatrix} 1+\xi_3 & \xi_1 \\ \xi_2 & 1-\xi_3 \end{bmatrix}.$$

The eigenvalues are

$$\lambda_{1,2} = 1 \pm \sqrt{\xi_3^2 + \xi_1\xi_2}.$$

At the origin, $\lambda_1 = \lambda_2$ is nondefective (with value 1) and we may take $P_1 = Q_1 = I$. Thus $B_k = A_k$, $k = 1, 2, 3$, and $\operatorname{tr} B_k = \operatorname{tr} A_k = 0$, so U, defined by (4.6), is $\frac{1}{2}I$. The spectral radius $\rho(x)$ is one at every point on the manifold where $\lambda_1 = \lambda_2$ is defective. Figure 4.1 shows a contour plot of $\rho(x)$ restricted to the (ξ_1, ξ_2) plane, where the defective manifold reduces to the coordinate axes.

COROLLARY. *There is always a direction d satisfying $f'(x_0; d) \leqq 0$, i.e., f never has a strongly unique local minimum at x_0.*

Proof. The proof is straightforward.

From both a practical and a theoretical point of view, obtaining a descent direction by making the active eigenvalue defective to first order is far from satisfactory. Because defective eigenvalues are very ill-conditioned, roundoff error may be overwhelming. Even in exact arithmetic, it is possible that a very small stepsize α may be required to make $f(x_0 + \alpha d) < f(x_0)$. In any case, finding the next descent direction to further reduce f may be very difficult, as explained in § 6. The following theorem greatly improves the situation.

THEOREM 4.4. *Suppose that $0 \notin \Phi$, i.e., f does not have a first-order local minimum at x_0 and therefore U, defined by (4.6), is not equal to $(1/t)I_t$. Then there exists a descent direction d along which $\lambda_1 = \cdots = \lambda_t$ is split into several nondefective eigenvalues. All eigenvalues maintain a common real part to first order, but they may have several different imaginary parts.*

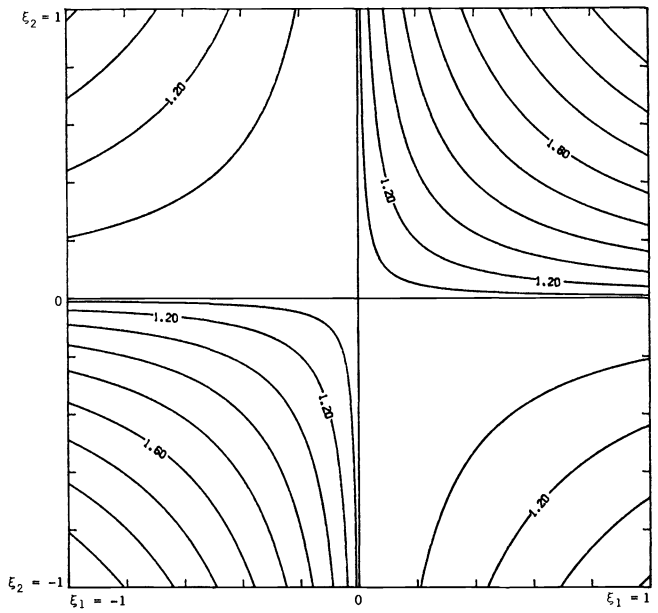


FIG. 4.1. *Contours of Example* 4.1 *in* $\xi_3 = 0$ *plane.*

Proof. Since $U \neq (1/t)I_t$, there exists a matrix E with *imaginary* eigenvalues such that $U:E \neq 0$, namely one of the following $t^2 - 1$ linearly independent matrices:

$$2e_p e_q^T - e_q e_p^T, \quad p,q = 1, \cdots, t, \quad p \neq q$$

or

$$e_p e_p^T - e_{p+1} e_{p+1}^T + 2e_p e_{p+1}^T - e_{p+1} e_p^T, \qquad p = 1, \cdots, t-1.$$

Now solve (4.8) for $[\delta_0, d^T]^T$, using the new right-hand side matrix E. As before, we obtain (4.9). Also as before, $f'(x_0; d)$ is the largest real part of the eigenvalues of $E - \delta_0 I_t$, i.e., $-\delta_0$, since E has imaginary eigenvalues. Thus a descent direction is obtained with the required property, since d may be replaced by $-d$ if $\delta_0 < 0$. Note that, to first order, multiplicity $t - 2$ is maintained along d, the common value being reduced by δ_0, while the other two eigenvalues split into a complex conjugate pair. It may be possible to split λ_1 further, with several eigenvalues taking on several different imaginary parts to first order, by choosing a less elementary matrix E with several different imaginary eigenvalues for the right-hand side of (4.8). □

The following question might arise: Can we obtain a descent direction along which $\lambda_1 = \cdots = \lambda_t$ is split into several distinct real eigenvalues? Obtaining such a descent direction d is much more difficult, since it is not true that $f'(x_0; -d) = -f'(x_0; d)$. If U has a negative real eigenvalue, such a descent direction may be obtained by using yz^T on the right-hand side of (4.8), where y^T, z are the left and right eigenvectors corresponding to the negative eigenvalue of U, as already explained. However, we have observed examples where there exists such a descent direction even if U has no negative eigenvalue.

Other examples have led us to the following conjecture that might be of interest.

CONJECTURE. Assume $n = 2$, $m = 3$, $\lambda_1(x_0) = \lambda_2(x_0)$ is nondefective, and $\{I_t, B_1, B_2, B_3\}$ has full rank. Then U has real eigenvalues if and only if there exist descent directions in *both* of the disconnected regions where $\lambda_{1,2}$ splits into a complex conjugate pair.

Remark. When there are descent directions in both of these disconnected regions, a convex combination of descent directions can give an ascent direction, namely in the region where $\lambda_{1,2}$ splits into a distinct real pair.

In the case $n = 2$, $t = 2$, $m = 3$, it is usually easy to find a descent direction by random search, since we need only that $\lambda_1(x_0)$ max Re $\mu_i < 0$, $i = 1, 2$. However, for larger t, the chance of finding a descent direction rapidly diminishes. In some randomly generated tests, we found that it was usually possible to obtain a descent direction with less than 500 random attempts for $n = t = 6$, $m = 35$, but this was not usually possible for $n = t = 8$, $m = 63$. Presumably the chance of success decreases exponentially with t.

We have now completed the discussion of case (2B). The degenerate case remains.

(2C) Nonsymmetric case, where neither (2A) nor (2B) applies.

This case generally applies if $m < c_N(t) = t^2 - 1$. As in case (1C), such situations are degenerate. Unlike the symmetric case, the nonsymmetric case no longer has an applicable optimality condition.

5. Optimality conditions in the case of several active nondefective multiple eigenvalues. Assume that $A(x_0)$ has several distinct active eigenvalues, all nondefective and with nonzero common modulus. Denote those that are real by λ_{jl}, $j = 1, \cdots, r$, and those that have positive imaginary parts by λ_{jl}, $j = r + 1, \cdots, s$, as described in the latter part of § 3. Recall that $\lambda_{j1} = \cdots = \lambda_{jt_j}$ is a multiple eigenvalue of multiplicity t_j, and recall the definition of $B_k^{(j)}$ given by (3.9). We now wish to generalize the results of the previous section.

(1) Symmetric case. This is easily generalized, since $r \leqq 2$ and $s = 0$. Details may be found in Overton (1988).

(2) Nonsymmetric case. To avoid confusing notation we entitle the three cases (A), (B), (C) somewhat differently than in § 4.

(2A) Nonsymmetric case, where we can obtain a descent direction without splitting a multiple eigenvalue or making it defective, or separating moduli.

For this case to apply, assume that the following linear system is solvable for $[\delta_1, \cdots, \delta_m, \varepsilon_{r+1}, \cdots, \varepsilon_{r+s}]^T \in \mathscr{R}^{m+s}$:

$$\sum_{k=1}^{m} \delta_k \lambda_{j1}(x_0) B_k^{(j)} = -I_{t_j}, \qquad j = 1, \cdots, r, \tag{5.1}$$

$$\sum_{k=1}^{m} \delta_k \operatorname{Re}(\bar{\lambda}_{j1}(x_0) B_k^{(j)}) = -I_{t_j}, \qquad j = r+1, \cdots, r+s, \tag{5.2}$$

$$\varepsilon_j I_{t_j} + \sum_{k=1}^{m} \delta_k \operatorname{Im}(\bar{\lambda}_{j1}(x_0) B_k^{(j)}) = 0, \qquad j = r+1, \cdots, r+s. \tag{5.3}$$

The system is generically solvable if $m > c_G(t_1, \cdots, t_{r+s})$, which is given by (2.5). Since we do not use the index i in this section, let $i = \sqrt{-1}$. Adding (5.2) to i times (5.3) we get

$$\sum_{k=1}^{m} \delta_k \bar{\lambda}_{j1}(x_0) B_k^{(j)} = -(1 + \varepsilon_j i) I_{t_j}, \qquad j = r+1, \cdots, r+s.$$

From Lemma 3.1, the first-order changes in the eigenvalue λ_{jl}, $l = 1, \cdots, t_j$, along the direction $d = \{\delta_1, \cdots, \delta_m\}^T$, are thus all the same quantity $-(1 + \varepsilon_j i)/\bar{\lambda}_{j1}(x_0)$, for each $j = r+1, \cdots, r+s$. Similarly, by (5.1), the first-order charges in λ_{jl}, $l = 1, \cdots, t_j$, are all $-1/\lambda_{j1}(x_0)$, for each $j = 1, \cdots, r$. Thus all multiple eigenvalues are preserved. Furthermore, by Lemma 3.4, or more specifically (3.11), the first-order change in $f_{jl} = \frac{1}{2}|\lambda_{jl}|^2$ is -1 for all $l = 1, \cdots, t_j$, $j = 1, \cdots, r+s$, i.e., all moduli are reduced along d and remain equal to first order.

(2B) Nonsymmetric case, where we can obtain a descent direction by splitting a multiple eigenvalue or making it defective or separating moduli, or else demonstrate optimality.

For this case to apply, assume that the coefficient matrix of the left-hand side of the following linear system has full column rank, and that the system is solvable. This case applies generically if $m = c_G(t_1, \cdots, t_{r+s})$. It also applies if $m > c_G(t_1, \cdots, t_{r+s})$, but f is minimized at x_0 on the manifold that preserves the nondefective multiplicities and the equal moduli. The linear system defines square dual matrices, $U_1, \cdots, U_{r+s}$, $V_{r+1}, \cdots, V_{r+s}$, of dimension $t_1, \cdots, t_{r+s}, t_{r+1}, \cdots, t_{r+s}$, respectively, by

$$\sum_{j=1}^{r} U_j : \lambda_{j1}(x_0) B_k^{(j)} + \sum_{j=r+1}^{r+s} U_j : \operatorname{Re}(\bar{\lambda}_{j1}(x_0) B_k^{(j)}) + \sum_{j=r+1}^{r+s} V_j : \operatorname{Im}(\bar{\lambda}_{j1}(x_0) B_k^{(j)}) = 0, \tag{5.4}$$

$$k = 1, \cdots, m,$$

$$\sum_{j=1}^{r+s} \operatorname{tr} U_j = 1, \tag{5.5}$$

$$\operatorname{tr} V_j = 0, \qquad j = r+1, \cdots, r+s. \tag{5.6}$$

The system (5.4)–(5.6) consists of $m + s + 1$ equations in $t_1^2 + \cdots + t_r^2 + 2t_{r+1}^2 + \cdots + 2t_{r+s}^2$ unknowns, so that it is square if $m = c_G(t_1, \cdots, t_{r+s})$.

THEOREM 5.1. *Define the dual matrices by* (5.4)–(5.6). *Then f has a first-order local minimum at x_0 if and only if $U_j = \kappa_j I_{t_j}$, where κ_j is a nonnegative real number, $j = 1, \cdots, r+s$, and $V_j = 0, j = r+1, \cdots, r+s$.*

Proof. First suppose that f does not have a first-order local minimum at x_0 and assume the given condition on the dual matrices holds. Let d be a descent direction for f from x_0. Then by Lemma 3.4,

$$\operatorname{Re}(\bar{\lambda}_{j1}(x_0)\mu_{jl}) < 0, \quad l = 1, \cdots, t_j, \quad j = 1, \cdots, r+s,$$

where $\{\mu_{jl}\}$ are the eigenvalues of $B^{(j)}(d)$, defined by (3.10). It follows that

$$\sum_{j=1}^{r+s} \kappa_j \operatorname{Re}(\operatorname{tr}(\bar{\lambda}_{j1}(x_0)B^{(j)}(d))) < 0.$$

(Note that $\sum_{j=0}^{r+s} \kappa_j t_j = 1$ by (5.5), so not all the $\{\kappa_j\}$ are zero.) Therefore, since the trace is the sum of diagonal elements,

$$\sum_{j=1}^{r+s} \kappa_j \operatorname{tr}(\operatorname{Re}(\bar{\lambda}_{j1}(x_0)B^{(j)}(d))) < 0.$$

But from (5.4), using the facts that $U_j = \kappa_j I_{t_j}$ and $V_j = 0$, and that $\bar{\lambda}_{j1}(x_0)B_k^{(j)}$ is real for $j = 1, \cdots, r$, we have

$$\sum_{j=1}^{r+s} \kappa_j \operatorname{tr}(\operatorname{Re}(\bar{\lambda}_{j1}(x_0)B_k^{(j)})) = 0, \qquad k = 1, \cdots, m.$$

By (3.10), this is a contradiction.

Now suppose that the given condition on the dual matrices does not hold. We wish to show that there exists a descent direction. Solve the following linear system in $[\delta_0, \delta_1, \cdots, \delta_m, \varepsilon_{r+1}, \cdots, \varepsilon_{r+s}] \in \mathscr{R}^{m+s+1}$:

$$\delta_0 I_{t_j} + \sum_{k=1}^{m} \delta_k \lambda_{j1}(x_0)B_k^{(j)} = E_j, \qquad j = 1, \cdots, r, \tag{5.7}$$

$$\delta_0 I_{t_j} + \sum_{k=1}^{m} \delta_k \operatorname{Re}(\bar{\lambda}_{j1}(x_0)B_k^{(j)}) = E_j, \qquad j = r+1, \cdots, r+s, \tag{5.8}$$

$$\varepsilon_j I_{t_j} + \sum_{k=1}^{m} \delta_k \operatorname{Im}(\bar{\lambda}_{j1}(x_0)B_k^{(j)}) = F_j, \qquad j = r+1, \cdots, r+s, \tag{5.9}$$

where the right-hand sides $\{E_j, F_j\}$ will now be defined. First note that the coefficient matrix of the left-hand side has full row rank, since it is the transpose of the coefficient matrix of the system (5.4)–(5.6), which defines the dual matrices. Now define all right-hand side matrices $\{E_j, F_j\}$ to be zero except one, namely E_h or F_h, which is to be defined by the first applicable case from the following list. At least one case must apply by assumption.

(i) Set $E_h = e_p e_q^T$ if there is a dual matrix U_h with a nonzero element in the (p, q) position, with $p \neq q$. Here e_p denotes the pth column of I_{t_h}.

(ii) Set $F_h = e_p e_q^T$ if there is a dual matrix V_h with a nonzero element in the (p, q) position, with $p \neq q$.

(iii) Set E_h to

$$e_p e_p^T - e_{p+1} e_{p+1}^T + e_p e_{p+1}^T - e_{p+1} e_p^T \tag{5.10}$$

if U_h is diagonal but has different pth and $(p+1)$th diagonal entries.

(iv) Set F_h to (5.10) if V_h is diagonal but has different pth and $(p+1)$th diagonal entries.

(v) The only other possibility is that $U_h = \kappa_h I$ where $\kappa_h < 0$ for some h, since we know $\operatorname{tr} V_j = 0, j = r+1, \cdots, r+s$ by (5.6). Set $E_h = -I_{t_h}$.

Now take inner products of $\{U_j\}$ with (5.7) and (5.8), respectively, and inner products of $\{V_j\}$ with (5.9), respectively. Summing the result and using (5.4) we obtain

$$\delta_0 \sum_{j=1}^{r+s} \operatorname{tr} U_j + \sum_{j=r+1}^{r+s} \varepsilon_j \operatorname{tr} V_j = U_h : E_h + V_h : F_h.$$

Here one of the terms on the right-hand side is zero. The other is nonzero by construction. Using (5.5), (5.6) we therefore have $\delta_0 \neq 0$, and, as before, we may take $\delta_0 > 0$ by reversing the sign of the right-hand side and the solution of (5.7)–(5.9). Now add i times (5.9) to (5.8) to obtain

$$\sum_{k=1}^{m} \delta_k \bar{\lambda}_{j1}(x_0) B_k^{(j)} = E_j + F_j - (\delta_0 + \varepsilon_j i) I_{t_j}, \qquad j = r+1, \cdots, r+s. \tag{5.11}$$

In cases (i)–(iv) the eigenvalues of all $E_j, j = 1, \cdots, r$, and all $E_j + F_j, j = r+1, \cdots, r+s$, are zero, even for $j = h$. Therefore, by (5.7) and (5.11),

$$\operatorname{Re}(\bar{\lambda}_{j1}(x_0)\mu_{jl}) = -\delta_0, \quad l = 1, \cdots, t_j, \quad j = 1, \cdots, r+s,$$

where $\{\mu_{jl}\}$ are the eigenvalues of

$$B^{(j)}(d) = \sum_{k=1}^{m} \delta_k B_k^{(j)}.$$

In case (v) the hth equation gives

$$\operatorname{Re}(\bar{\lambda}_{h1}(x_0)\mu_{hl}) = -\delta_0 - 1$$

since $E_h = -I$. In both cases $f'(x_0; d) < 0$, where $d = [\delta_1, \cdots, \delta_m]^T$, by Lemma 3.4. □

Remark. In cases (i)–(iv), descent is obtained by maintaining all eigenvalue multiplicities but making $\lambda_{h1} = \cdots = \lambda_{ht_h}$ defective (to first order). We could just as well split $\lambda_{h1} = \cdots = \lambda_{ht_h}$ so that the change in all eigenvalues in the group has a common positive component in the direction (in the complex plane) $-\lambda_{h1}(x_0)$, and has different components in the orthogonal direction, i.e., tangent to the circle centered at the origin and passing through $\lambda_{h1}(x_0)$. This is what we did in Theorem 4.4, where the multiple eigenvalue is real. All we need do is set E_h or F_h, respectively, to a matrix with imaginary eigenvalues and nonzero inner product with U_h or V_h. In case (v), descent is obtained by preserving all nondefective eigenvalue multiplicities but reducing the modulus of λ_{hl} by more than the moduli of the other eigenvalues.

Remark. In the case $s = 0$, $t_j = 1$, $j = 1, \cdots, r$, Theorem 5.1 reduces to the standard min-max optimality condition where only case (v) applies. In the case $s = 0$, $r = 1$, the theorem reduces to Theorem 4.3. In the case $r = 0$, $s = 1$, the theorem reduces to a statement about splitting a multiple eigenvalue which is one of a single active complex conjugate pair.

We conclude this section with two examples.

Example 5.1. Reconsider Example 1.1. At $x_0 = [-1]$, we have $r = 2$, $t_1 = t_2 = 1$, $s = 0$. The codimension of the manifold defined by $|\lambda_1(x)| = |\lambda_2(x)|$ is $c_G(1, 1; 0) = 1$, so since $m = 1$, the dimension of the manifold is zero. The optimality condition is checked as follows. We have

$$\lambda_1(x_0) = 1, \qquad \lambda_2(x_0) = -1,$$

$$[Q_1\ Q_2] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad [P_1\ P_2] = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

$$B_1^{(1)} = \begin{bmatrix} \frac{1}{2} \end{bmatrix}, \qquad B_1^{(2)} = \begin{bmatrix} \frac{3}{2} \end{bmatrix}.$$

Equations (5.4)–(5.6), which define the dual matrices, in this case scalars, give

$$\frac{1}{2}U_1 - \frac{3}{2}U_2 = 0, \qquad U_1 + U_2 = 1.$$

The solution is $U_1 = \kappa_1 = \frac{3}{4}$, $U_2 = \kappa_2 = \frac{1}{4}$, so x_0 is indeed optimal.

Example 5.2. Let $n = 10$, let $x_0 = [0, \cdots, 0]^T$, and define $A_0 = A(x_0)$ by

$$A_0 = \begin{bmatrix} \sqrt{2} & & & & & & & & & \\ & \sqrt{2} & & & & & & & & \\ & & \sqrt{2} & & & & & & & \\ & & & \sqrt{2} & & & & & & \\ & & & & 1 & -1 & & & & \\ & & & & 1 & 1 & & & & \\ & & & & & & 1 & -1 & & \\ & & & & & & 1 & 1 & & \\ & & & & & & & & 1 & \\ & & & & & & & & & 0 \end{bmatrix}.$$

This matrix has one active quadruple real eigenvalue and one active double complex conjugate pair of eigenvalues, all with modulus $\sqrt{2}$. Thus $r = 1$, $s = 1$, $t_1 = 4$, $t_2 = 2$. In order for a generic family $A(x)$ to have x_0, and only x_0, as a point where $A(x)$ has a quadruple real eigenvalue and a complex conjugate pair with the same modulus, we require

$$m = c_G(4; 2) = 16 + 8 - 1 - 1 = 22.$$

The component matrices $\{A_k\}$, $k = 1, \cdots, 22$, are randomly generated by setting the elements, in the order $(A_1)_{11}, (A_1)_{1,2}, \cdots, (A_1)_{1,n}, \cdots, (A_1)_{n,n}, (A_2)_{1,1}, \cdots, (A_m)_{n,n}$, to the sequence ψ_ν, $\nu = 1, 2, \cdots$, defined by

$$\psi_\nu = \frac{\theta_\nu}{4095}, \qquad \theta_\nu = (445\theta_{\nu-1} + 1) \bmod 4096$$

and $\theta_0 = 1$.

We have $\lambda_{1,l} = \sqrt{2}$, $l = 1, \cdots, 4$, $\lambda_{2,l} = 1 + i$, $l = 1, 2$, and

$$P_1 = Q_1 = e_1e_1^T + e_2e_2^T + e_3e_3^T + e_4e_4^T,$$

$$P_2 = -\frac{i}{\sqrt{2}}(e_5e_1^T + e_7e_2^T) + \frac{1}{\sqrt{2}}(e_6e_1^T + e_8e_2^T),$$

$$Q_2 = \frac{i}{\sqrt{2}}(e_5e_1^T + e_7e_2^T) + \frac{1}{\sqrt{2}}(e_6e_1^T + e_8e_2^T).$$

Here e_p is the pth column of the identity matrix of the appropriate dimension, so that P_1, Q_1 are 10×4 and P_2, Q_2 are 10×2. Forming the system (5.4)–(5.6) and solving it, we obtain

$$U_1 = \begin{bmatrix} .455 & .040 & .039 & -.231 \\ -.110 & .094 & -.057 & -.187 \\ .007 & .017 & .335 & -.043 \\ -.227 & .092 & .002 & -.187 \end{bmatrix},$$

$$U_2 = \begin{bmatrix} .354 & -.017 \\ -.114 & -.050 \end{bmatrix}, \qquad V_2 = \begin{bmatrix} -.515 & -.078 \\ .338 & .515 \end{bmatrix}.$$

Thus there are many possible descent directions. For example, we have the following:

(i) Let $E_1 = -e_4 e_1^T$, $E_2 = 0$, $F_2 = 0$. Solving (5.7)–(5.9) we obtain $(\delta_0, d, \varepsilon_2)$ with $f'(0; d) = -\delta_0 = -.227$. Along this direction $\lambda_{1,1} = \cdots = \lambda_{1,4}$ does not split but becomes defective (to first order).

(ii) Let $E_1 = -e_2 e_1^T + e_1 e_2^T$, $E_2 = 0$, $F_2 = 0$. We get $f'(0; d) = -\delta_0 = -.150$. Because E_1 has imaginary eigenvalues, $\lambda_{1,1} = \cdots = \lambda_{1,4}$ splits into a complex conjugate pair and a double real eigenvalue (to first order).

(iii) Let $E_1 = 0$, $E_2 = -e_2 e_1^T$, $F_2 = 0$. We get $f'(0; d) = -\delta_0 = -.114$. This time it is the double complex conjugate pair of eigenvalues that becomes defective (to first order).

(iv) Let $E_1 = 0$, $E_2 = -e_2 e_1^T + e_1 e_2^T$, $F_2 = 0$. We get $f'(0; d) = -\delta_0 = -.097$. The double complex conjugate pair of eigenvalues splits in directions tangent to the circle in the complex plane centered at the origin with radius $\sqrt{2}$.

Finally, there is the degenerate case.

(2C) Nonsymmetric case, where neither (2A) nor (2B) applies.

This case generally applies if $m < c_G(t_1, \cdots, t_{r+s})$. As before such situations are degenerate, and the optimality condition does not apply.

6. The defective case. If $A(x_0)$ has a defective active eigenvalue, none of the previous results apply. In such cases it seems very hard to determine in general whether x_0 is a local minimizer of f, and, if not, to generate a descent direction. Indeed, it is well known that even determining the Jordan structure of $A(x_0)$ is difficult numerically.

Suppose there is one real active multiple eigenvalue $\lambda_1(x_0) = \cdots = \lambda_t(x_0)$, and suppose the orders of the corresponding Jordan blocks are $u_1 \geqq \cdots \geqq u_p$, $1 \leqq p \leqq t$. The codimension of the manifold on which the same Jordan structure is preserved is $c = c_D(u_1, \cdots, u_p)$, given by (2.6). If $m > c$, then generically the dimension of this manifold is at least one, and if x_0 does not minimize f on the manifold, it seems reasonable to suppose that a descent direction exists. This is not clear, however, since f is not Lipschitz along lines through x_0.

If $m = c$, then generically x_0 is the only point where $A(x)$ has the given Jordan structure. If $\lambda_1(x_0)$ is derogatory, i.e., there is more than one Jordan block corresponding to $\lambda_1(x_0)$, it may be possible to decrease $f(x)$ by making $\lambda_1(x)$ "more defective," i.e., moving to a point x where two of the Jordan blocks combine to form a larger block. Such points lie on a manifold with smaller codimension and hence larger dimension. If $m = c$ and $\lambda_1(x_0)$ is nonderogatory, i.e., $p = 1$, it will generally be necessary to split the multiple eigenvalue to obtain a reduction in f. It seems that the cases where x_0 is most likely to be a minimum are where $\lambda_1(x_0)$ is nonderogatory.

If $\lambda_1(x_0)$ is nonderogatory, an arbitrary perturbation of x with size ε will generally perturb the eigenvalues by $O(\varepsilon^{1/t})$. More specifically, the eigenvalues can be expanded

in Puiseux series; see Kato (1984, p. 65). The sum of the perturbed eigenvalues is analytic in ε (Kato (1984, p. 78)); accordingly, the $O(\varepsilon^{1/t})$ changes in the t eigenvalues are generally of equal magnitude and along directions in the complex plane separated by angles of $2\pi/t$. It follows that if $t > 2$, the spectral radius is increased by $O(\varepsilon^{1/t})$. If $t = 2$, the only case in which the spectral radius changes by $O(\varepsilon)$ is that in which the eigenvalues split into a complex conjugate pair; or more generally, if $\lambda_1(x_0)$ is complex, that in which the changes in the eigenvalues are tangent to the circle in the complex plane centred at the origin and passing through $\lambda_1(x_0)$. However, it is also true in the nondefective case that arbitrary perturbations to x generally increase the spectral radius; the question is whether a properly chosen perturbation can decrease f. It may be possible, even in the nonderogatory case, to perturb x so that the spectral radius is decreased. This would require that the first nonzero term in the Puiseux series be either an imaginary term of size $O(\varepsilon^{1/2})$ or a real term of size $O(\varepsilon)$. It might be achieved, for example, by splitting off a complex conjugate pair of eigenvalues and preserving multiplicity $t - 2$.

Consider Example 1.1. At $x_0 = 0$, $\lambda_1(x_0)$ is defective, with $n = t = 2$, $p = 1$. We have $c = m = 1$, and, indeed, x_0 is the only point where $\lambda_1(x)$ is defective. The point x_0 is a local minimizer of f. Now generalize the example to

$$A(x) = \begin{bmatrix} 1 + \gamma\xi_1 & 1 \\ -\xi_1 & 1 + \gamma\xi_1 \end{bmatrix}$$

with $x_0 = [0]$. Regardless of γ, the eigenvalues of $A(x)$ are real for $\xi_1 < 0$ and we may legitimately generalize the notion of directional derivative to say that $f'(0; -1) = +\infty$. For $\xi_1 > 0$, the eigenvalues are a complex conjugate pair, with

$$\lambda_{1,2}(\xi_1) = 1 + \gamma\xi_1 \pm i\sqrt{\xi_1}$$

so that

$$f'(0; +1) = \gamma + \tfrac{1}{2}.$$

Thus zero is a first-order local minimizer if and only if $\gamma \geqq -\frac{1}{2}$. In fact, we may without difficulty extend the definition of Clarke's generalized gradient to handle the case $m = 1$ regardless of whether $\lambda_1(x_0)$ is defective. In this particular case we obtain

$$\partial f(0) = [-\infty, \gamma + \tfrac{1}{2}]$$

so that, for any γ, f has a first-order local minimum at zero if and only if $0 \in \partial f(0)$.

The reason that duality theory, particularly the theorems in §§ 4 and 5, is so useful is that information computed only at x_0 defines dual variables, in our case matrices, that resolve the question of optimality and give information regarding descent directions. If $\lambda_1(x_0)$ is defective, however, it does not seem possible, even in the simple case just described, to resolve optimality directly from the information given by the Jordan form of $A(x_0)$ together with the component matrices $\{A_k\}$. It is possible, of course, to determine whether a given direction d is a descent direction by looking at the limit of the well defined quantities $f'(x_0 + \varepsilon d; d)$, where $\varepsilon > 0$ and $A(x_0 + \varepsilon d)$ has distinct eigenvalues, but this is of little use when $m > 1$.

Let us turn to Example 1.2 (see Fig. 1.2). We see that at, say, $x_0 = [1, 0]^T$, it is not trivial to determine which directions into the "complex region" are descent directions. In this case, the defective manifold shown in Fig. 1.2 is linear, so reducing f by keeping the eigenvalue defective poses no difficulty.

Finally, consider the following example.

Example 6.1. Let $n = 3$, $m = 2$ and define

$$A(x) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} + \xi_1 \begin{bmatrix} .5 & -.2 & -.4 \\ .7 & 1.2 & 1 \\ -2 & .8 & -.3 \end{bmatrix} + \xi_2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Let $x_0 = [0, 0]^T$. At x_0, A has a nonderogatory triple eigenvalue. The codimension $c = 3 - 1 = 2$. Since $m = 2$, x_0 is the only point with this Jordan structure. Figure 6.1 gives a contour plot of $\rho(\xi_1, \xi_2)$. Figure 6.2 shows graphs of $\rho(\xi_1, \xi_2)$ along the lines $\xi_2 = 0.1$, $\xi_2 = 0$ and $\xi_2 = -0.1$, respectively.

There is a curve clearly visible in Fig. 6.1 across which $\rho(x)$ is not differentiable. Along the part of the curve above the point x_0, $A(x)$ is defective; more specifically, the triple eigenvalue splits into one defective double real eigenvalue and one single eigenvalue. On the part of the curve below x_0, $A(x)$ is not defective, and in fact it has distinct eigenvalues, one complex conjugate pair and one real eigenvalue. Along this part of the curve, $\rho(x)$ is a Lipschitz max function, with the complex conjugate pair and the real eigenvalue achieving the same modulus. Theorem 5.1 is trivially applicable at these points. It can be seen that ρ is Lipschitz along $\xi_2 = -0.1$ (Fig. 6.2(c)), that ρ is not Lipschitz along $\xi_2 = 0.1$ (Fig. 6.2(a)), and that ρ has even more rapid variation along $\xi_2 = 0$ (Fig. 6.2(b)); this is because a triple eigenvalue is being perturbed in the last case. There is another curve emanating up from x_0 along which the triple eigenvalue also splits into one defective double real eigenvalue and one single eigenvalue. This curve is not visible in the contour plot, since it is the *distinct* eigenvalue that has the maximum modulus. Thus the "defective manifold" has a cusp at x_0. This is consistent with the illustration given by Arnold (1971, p. 38); the manifold here corresponds to a cross-section of the one shown by Arnold.

We note that ρ is apparently locally but not globally minimized at x_0. There are lower values of ρ on the curve of discontinuity towards the bottom of Fig. 6.1.

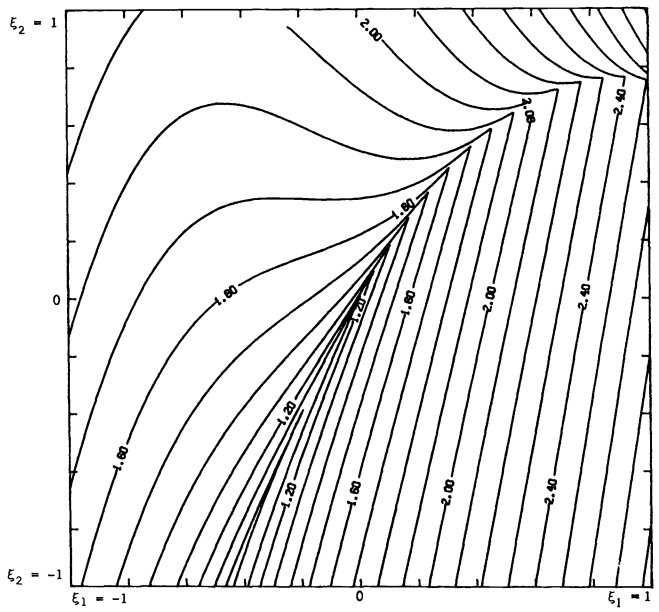


FIG. 6.1. *Contour plot of Example* 6.1.

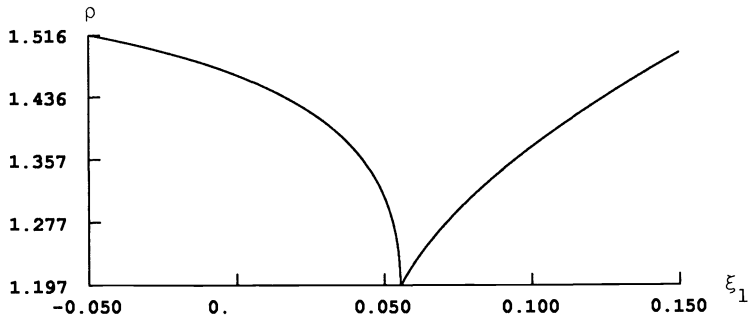


FIG. 6.2(a)

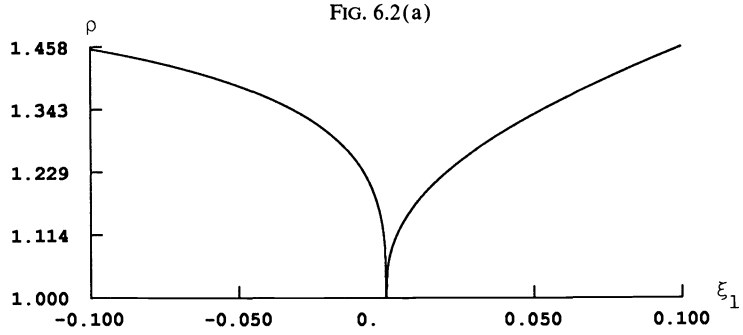


FIG. 6.2(b)

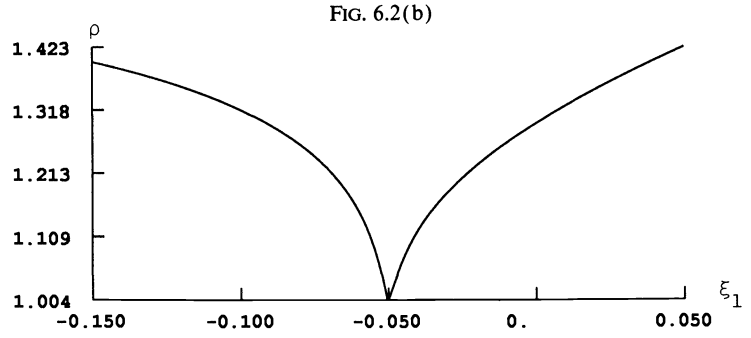


FIG. 6.2(c)

In summary, the question of optimality seems very hard to resolve in the defective case, and many interesting questions remain open.

Acknowledgments. The authors thank Dr. M. R. Osborne for many stimulating discussions, and Dr. Alastair Spence for bringing the work of Arnold to their attention. The first author thanks his hosts at the Australian National University, especially Mike Osborne, for their warm hospitality.

REFERENCES

V. I. ARNOLD (1971), *On matrices depending on parameters*, Russian Math. Surveys, 26, No. 2 pp. 29–43.

——— (1983), *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, Berlin, New York.

S. BOYD (1988), *Structured and simultaneous Lyapunov functions for system stability problems*, Information Systems Laboratory Report L-104-88-1, Stanford University, Stanford, CA.

J. V. BURKE (1987), private communication.

F. H. CLARKE (1975), *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205, pp. 247–262.

——— (1983), *Optimization and Nonsmooth Analysis*, Wiley–Interscience, New York.

J. W. DEMMEL (1983), *A Numerical Analyst's Jordan Canonical Form*, Ph.D. thesis, Computer Science Dept., University of California, Berkeley, CA.

R. FLETCHER (1981), *Practical Methods of Optimization*, Vol. 2, John Wiley, Chichester, New York.

——— (1985), *Semi-definite matrix constraints in optimization*, SIAM J. Control Optim., 23, pp. 493–513.

S. FRIEDLAND (1978), *Extremal eigenvalue problems*, Bol. Soc. Brasil. Mat., 9, pp. 13–40.

S. FRIEDLAND, J. NOCEDAL, AND M. L. OVERTON (1987), *The formulation and analysis of numerical methods for inverse eigenvalue problems*, SIAM J. Numer. Anal., 24, pp. 634–667.

G. H. GOLUB AND C. VAN LOAN (1983), *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD.

V. A. KAMENETSKII AND E. S. PYATNITSKII (1987), *Gradient method of constructing Lyapunov functions in problems of absolute stability*, Automat. Remote Control, 48, pp. 1–8.

T. KATO (1984), *Perturbation Theory for Linear Operators*, 2nd edition, Springer-Verlag, Berlin, New York.

W. LEDERMANN (1937), *On the rank of the reduced correlational matrix in multiple-factor analysis*, Psychometrika, 2, pp. 85–93.

P. M. MÄKILÄ AND H. T. TOIVONEN (1987), *Computational methods for parametric LQ Problems—a survey*, IEEE Trans. Automat. Control, AC-32, pp. 658–671.

L. F. MILLER, R. G. COCHRAN, AND J. W. HOWZE (1978), *Output feedback stabilization by minimization of a spectral radius functional*, Internat. J. Control, 27, pp. 455–462.

P. NOWASAD (1968), *Isoperimetric eigenvalue problems in algebras*, Comm. Pure Appl. Math., 21, pp. 401–465.

M. L. OVERTON (1988), *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM J. Matrix Anal. Appl., 9, pp. 256–268.

R. T. ROCKAFELLAR (1970), *Convex Analysis*, Princeton University Press, Princeton, NJ.

——— (1981), *The Theory of Subgradients and Its Application to Problems of Optimization: Convex and Nonconvex Functions*, in Research and Education in Mathematics 1, Heldermann-Verlag, Berlin.

——— (1985), *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal. Theory Methods Appl., 9, pp. 665–698.

J. VON NEUMANN AND E. WIGNER (1929), *Über das Verhalten von Eigenwerten bei adiabatischen Prozessen*, Physik. Zeitschr., 30, pp. 467–470.

D. M. YOUNG (1971), *Iterative Solution of Large Linear Systems*, Academic Press, New York.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 4, October 1988

004

THE PERIODIC LYAPUNOV EQUATION*

PAOLO BOLZERN† AND PATRIZIO COLANERI‡

Abstract. This paper presents an overview of the periodic Lyapunov equation, both in discrete time and in continuous time. Together with some selected results that have recently appeared in the literature, the paper provides necessary and sufficient conditions for the existence and uniqueness of periodic solutions.

Key words. linear periodic systems, Lyapunov equation, inertia of matrices

AMS(MOS) subject classifications. 34A30, 39A10, 34C25, 93D05

1. Introduction. The Lyapunov equation arises in a large variety of problems in linear systems theory. In particular, it represents a basic tool for stability analysis of linear systems and for state covariance computation in a stochastic framework. It is also useful in the analysis of the Riccati equations encountered in optimal filtering and control problems.

Due to its importance, the Lyapunov equation has long deserved considerable attention, starting with the pioneering work of Lyapunov himself [1]. Since then, most results have concerned the time-invariant case (algebraic Lyapunov equation). In particular, in [2], [3], the so-called "Lyapunov lemma" was established, linking the existence of a positive-definite solution to the asymptotic stability of the underlying system. Other authors (see, e.g., [4], [5]) investigated spectral properties of solutions, which led to further developments known as the "inertia theory."

In the present paper, we consider the periodic Lyapunov equations, namely the discrete-time periodic Lyapunov equation (DPLE)

$$P(t+1)=A(t)P(t)A(t)'+B(t)B(t)' \tag{1}$$

where $t \in Z$, and $A(\cdot)\colon Z \to R^{n\times n}$, $B(\cdot)\colon Z \to R^{n\times m}$ are periodic matrices of period $T \in Z^+$, and the continuous-time periodic Lyapunov equation (CPLE)

$$\dot{P}(t)=A(t)P(t)+P(t)A(t)'+B(t)B(t)' \tag{2}$$

where $t \in R$, and $A(\cdot)\colon R \to R^{n\times n}$, $B(\cdot)\colon R \to R^{n\times m}$ are continuous periodic matrices of period $T \in R^+$.

In many problems, the dual versions of these equations must be considered. However, we will concentrate only on the DPLE and CPLE above, since the analysis of the dual equations can be carried out by standard duality considerations.

Some results obtained in the algebraic case have been recently extended to the periodic Lyapunov equations. In particular, the "periodic Lyapunov lemma" was worked out in [6], [7], while the "periodic inertia theory" was addressed in [8]–[11]. Results concerning the Lyapunov equations with generally time-varying coefficients can be found in [12] and [13].

Despite such a large amount of research, a complete picture of necessary and sufficient conditions for the existence and uniqueness of periodic solutions of the DPLE and CPLE

* Received by the editors June 1, 1987; accepted for publication (in revised form) March 1, 1988. This work was supported by Centro di Teoria dei Sistemi del Consiglio Nazionale delle Ricerche and by Ministero della Pubblica Istruzione.

† Dipartimento di Elettronica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy.

‡ Centro di Teoria dei Sistemi del Consiglio Nazionale delle Ricerche and Dipartimento di Elettronica, Politecnico di Milano, Piazza Leonardo da Vinci, 20133 Milan, Italy.

is still lacking. Filling this deficiency is the main purpose of this paper. However, a few previous results are also reported herein, for the sake of completeness. The implications with the newly developed theorems will be appropriately pointed out.

Basically, criteria for the existence and uniqueness of real symmetric (possibly positive (semi-)definite) T-periodic solutions of the DPLE and CPLE are provided. The derivation hinges on reducing to a suitable discrete-time algebraic Lyapunov equation (DALE), the solutions of which are shown to be the periodic generators of the DPLE and CPLE.

Major attention will be devoted to the DPLE. Actually, the analysis of the CPLE is completely analogous and leads to conclusions that are (mostly) formally identical.

The main results of this paper are stated in terms of the structural properties and the canonical decomposition of linear periodic systems. The reader who is unfamiliar with these topics is referred to [14] for an exhaustive survey.

The paper is organized as follows. After a brief review of some basic concepts regarding discrete-time linear periodic systems (§ 2), the relationship between the DPLE and a suitable DALE is discussed in § 3. In § 4, conditions for the existence and uniqueness of the solutions of the DALE are worked out. Such conditions are extended to the DPLE in § 5. Section 6 is devoted to the inertia theory for the DPLE. Finally, the continuous-time case (CPLE) is briefly treated in § 7.

2. Discrete-time linear periodic systems—basic concepts. Consider the system

$$x(t+1)=A(t)x(t)+B(t)u(t) \tag{3}$$

where $t \in Z$, and $A(\cdot)\colon Z \to R^{n\times n}$, $B(\cdot)\colon Z \to R^{n\times m}$ are periodic matrices of period $T \in Z^+$. The transition matrix over $[\tau, t]$ associated with $A(\cdot)$ will be denoted by $\Phi(t, \tau)$, i.e.,

$$\Phi(t,\tau)=\begin{cases} A(t-1)A(t-2)\cdots A(\tau), & t>\tau, \\ I_n, & t=\tau. \end{cases}$$

The monodromy matrix of $A(\cdot)$ at time τ is defined as $\bar{\Phi}_\tau = \Phi(\tau + T, \tau)$. It is well known [14] that the eigenvalues of $\bar{\Phi}_\tau$ are independent of τ, and that system (3) is asymptotically stable if and only if all the eigenvalues of $\bar{\Phi}_\tau$ lie inside the open unit disk in the complex plane. In this case, we will say, for short, that $A(\cdot)$ is asymptotically stable. Moreover, denote by $W(\tau, t)$ the reachability Gramian matrix associated with $(A(\cdot), B(\cdot))$, i.e.,

$$W(\tau,t)=\sum_{j=\tau+1}^{t}\Phi(t,j)B(j-1)B(j-1)'\Phi(t,j)'.$$

Also, let $\bar{W}_\tau = W(\tau, \tau + T)$ represent the single-period reachability Gramian, and $\bar{D}_\tau$ be any matrix such that $\bar{D}_\tau\bar{D}_\tau' = \bar{W}_\tau$.

The structural properties of system (3) (reachability, controllability, and stabilizability of the pair $(A(\cdot), B(\cdot))$) have been analyzed in several papers (see, e.g., [14] and the references quoted therein). It is worth reminding the reader that reachability refers to the possibility of reaching any state from the origin, while controllability refers to the possibility of driving any initial state to zero (see, e.g., [15]). In the present paper, we concentrate only on the characterizations stated in terms of the constant pair $(\bar{\Phi}_\tau, \bar{D}_\tau)$, where τ is fixed. Precisely, the following criteria hold (see [16]).

PROPOSITION 1. *The pair $(A(\cdot), B(\cdot))$ is reachable at τ if and only if $(\bar{\Phi}_\tau, \bar{D}_\tau)$ is reachable.*

PROPOSITION 2. *The pair $(A(\cdot), B(\cdot))$ is controllable if and only if $(\bar{\Phi}_\tau, \bar{D}_\tau)$ is controllable.*

PROPOSITION 3. *The pair* $(A(\cdot), B(\cdot))$ *is stabilizable if and only if* $(\bar{\Phi}_\tau, \bar{D}_\tau)$ *is stabilizable.*

Notice that reachability of the periodic pair $(A(\cdot), B(\cdot))$ at τ does not imply in general reachability at a different time point. Now, let $X_r(t)$ and $X_c(t)$ denote the reachability and controllability subspaces at time t of system (3), and let $Z_{\tau r}$ and $Z_{\tau c}$ denote the reachability and controllability subspaces of $(\bar{\Phi}_\tau, \bar{D}_\tau)$. Then the following relationship can be proven.

PROPOSITION 4.

$$X_r(\tau) = Z_{\tau r}, \qquad X_c(\tau) = Z_{\tau c}.$$

The conclusions of Proposition 4 can be easily derived from the results given in [17].

It can be shown that the dimension of $X_c(\tau)$ does not vary with τ (see [14]). This enables us to perform the Kalman canonical decomposition of system (3) into the controllable and uncontrollable parts. Specifically, a nonsingular T-periodic state-space transformation exists that puts the matrices $A(\cdot)$ and $B(\cdot)$ in the canonical form

$$A(\cdot) = \begin{bmatrix} A_1(\cdot) & A_2(\cdot) \\ 0 & A_3(\cdot) \end{bmatrix}, \qquad B(\cdot) = \begin{bmatrix} B_1(\cdot) \\ 0 \end{bmatrix}$$

where the pair $(A_1(\cdot), B_1(\cdot))$ is controllable. According to such a decomposition, matrices $\bar{\Phi}_\tau$ and $\bar{D}_\tau$ take on the following form:

$$\bar{\Phi}_\tau = \begin{bmatrix} \Phi_1 & \Phi_2 \\ 0 & \Phi_3 \end{bmatrix}, \qquad \bar{D}_\tau = \begin{bmatrix} D_1 \\ 0 \end{bmatrix}.$$

It is straightforward to verify that this partition coincides with the standard canonical decomposition of the constant pair $(\bar{\Phi}_\tau, \bar{D}_\tau)$. In particular, Φ_1 represents both the controllable part of $(\bar{\Phi}_\tau, \bar{D}_\tau)$ and the monodromy matrix of the controllable part $A_1(\cdot)$ of $(A(\cdot), B(\cdot))$.

3. Time-invariant reformulation of the DPLE. In this section, we will point out the relationship between the periodic solutions of the DPLE and the constant solutions of a suitable discrete-time algebraic Lyapunov equation (DALE). We will be concerned with real symmetric T-periodic solutions of the DPLE. Hence, we will often omit the phrase "real symmetric," for simplicity.

By solving recursively the DPLE starting from $P(\tau) = P_\tau$, the solution at $t > \tau$ is given by

$$P(t) = \Phi(t,\tau) P_\tau \Phi(t,\tau)' + W(\tau, t).$$

Since we are looking for periodic solutions of the DPLE ($P(\tau + T) = P(\tau)$ for all τ), the periodic generator P_τ must satisfy the following algebraic Lyapunov equation (DALE):

$$P_\tau = \bar{\Phi}_\tau P_\tau \bar{\Phi}_\tau' + \bar{D}_\tau \bar{D}_\tau'. \tag{4}$$

Thus, a bijective correspondence between the solutions $\bar{P}_\tau$ of (4) and the T-periodic solutions $\tilde{P}(\cdot)$ of the DPLE can be established. In particular, the following propositions hold.

PROPOSITION 5. *The* DPLE (1) *admits a T-periodic solution* $\tilde{P}(\cdot)$ *if and only if the* DALE (4) *admits a solution* $\bar{P}_\tau$.

PROPOSITION 6. *The* DPLE (1) *admits a unique T-periodic solution* $\tilde{P}(\cdot)$ *if and only if the* DALE (4) *admits a unique solution* $\bar{P}_\tau$.

In many cases, we are interested in the T-periodic solutions of the DPLE (1) that are positive definite or semidefinite. In this respect, the following results can be established.

PROPOSITION 7. *The* DPLE (1) *admits a T-periodic solution* $\tilde{P}(\cdot)$ *that is positive semidefinite at any* $t(\tilde{P}(t) \geqq 0$, *for all* t) *if and only if the* DALE (4) *admits a positive-semidefinite solution* $\bar{P}_\tau$.

PROPOSITION 8. *The* DPLE (1) *admits a T-periodic solution* $\tilde{P}(\cdot)$ *that is positive definite at time* τ *if and only if the* DALE (4) *admits a positive-definite solution* $\bar{P}_\tau$.

Notice that a T-periodic solution $\tilde{P}(\cdot)$ that is positive definite at τ is also positive definite at any t, if $A(t)$ is nonsingular for all t. In general, it is only positive semidefinite.

Remark. The DALE (4) can also be seen as the Lyapunov equation associated with a time-invariant sampled-state representation of the periodic system (3), i.e.,

$$z_\tau(k+1) = \bar{\Phi}_\tau z_\tau(k) + \bar{D}_\tau v_\tau(k), \qquad k \in Z$$

where

$$z_\tau(k) = x(\tau + kT),$$

$$v_\tau(k) = [u(\tau + kT)'u(\tau + 1 + kT)' \cdots u(\tau + (k+1)T - 1)']'$$

and $\bar{D}_\tau$ is the following factor of $\bar{W}_\tau$:

$$\bar{D}_\tau = [\Phi(\tau + T, \tau + 1)B(\tau)\Phi(\tau + T, \tau + 2)B(\tau + 1) \cdots B(\tau + T - 1)].$$

This useful reformulation was introduced in [18].

4. Existence and uniqueness conditions for the solutions of the DALE. In view of Propositions 5–8, our attention must focus on the properties of the constant solutions of the DALE (4). Thus, consider a discrete-time algebraic Lyapunov equation (DALE):

$$Q = FQF' + GG' \tag{5}$$

where $F \in R^{n \times n}$, $G \in R^{n \times m}$, and the unknown Q is real and symmetric.

The purpose of this section is to derive a set of necessary and sufficient conditions for the existence and uniqueness of the solutions of the DALE (5).

Remark. Consider a nonsingular $n \times n$ matrix S and the DALE

$$\hat{Q} = \hat{F}\hat{Q}\hat{F}' + \hat{G}\hat{G}' \tag{6}$$

where $\hat{F} = SFS^{-1}$ and $\hat{G} = SG$. It is easy to see that $\bar{Q}$ is a solution of (5) if and only if $\hat{\bar{Q}} = S\bar{Q}S'$ is a solution of (6). Moreover, such a correspondence preserves symmetry and positive (semi-)definiteness of the solution. Hence, there is no loss of generality in considering, when necessary, suitable canonical forms of the pair (F, G).

A major role in the sequel will be played by the Kalman canonical decomposition of (F, G) into the controllable and uncontrollable parts, i.e.,

$$\hat{F} = \begin{bmatrix} F_1 & F_2 \\ 0 & F_3 \end{bmatrix}, \qquad \hat{G} = \begin{bmatrix} G_1 \\ 0 \end{bmatrix} \tag{7}$$

where (F_1, G_1) is controllable. F_1 and F_3 are the controllable and uncontrollable parts of (F, G), respectively.

The results concerning the solutions of the DALE (5) are summarized in the following theorems, whose proofs are given in the Appendix. The symbol $\mu_i(M)$ will indicate the ith eigenvalue of the square matrix M.

THEOREM 1. *The* DALE (5) *admits a unique solution* $\bar{Q}$ *if and only if* F *does not have reciprocal eigenvalues* $(\mu_i(F)\mu_j(F) \neq 1$ *for all* $i, j)$.

THEOREM 2. *The* DALE (5) *admits a solution Q if and only if, for each $\mu \in C$, $y \in C^n$, $z \in C^n$ such that $F'y = \mu y$ and $F'z = \mu^{-1}z$, it results in $y^*GG'z = 0$ (where * denotes conjugate transpose).*

THEOREM 3. *The* DALE (5) *admits a positive definite solution $\bar{Q}$ if and only if*

(i) *F_1 is asymptotically stable ($|\mu_i(F_1)| < 1$, for all i);*

(ii) *$|\mu_i(F_3)| = 1$ for all i;*

(iii) *F_3 is diagonalizable;*

(iv) *(F_1, G_1) is reachable.*

THEOREM 4. *The* DALE (5) *admits a positive semidefinite solution $\bar{Q}$ if and only if F_1 is asymptotically stable ($|\mu_i(F_1)| < 1$, for all i).*

THEOREM 5. *The* DALE (5) *admits a unique positive-definite solution $\bar{Q}$ if and only if:*

(i) *F is asymptotically stable ($|\mu_i(F)| < 1$, for all i);*

(ii) *(F, G) is reachable.*

THEOREM 6. *The* DALE (5) *admits a unique positive-semidefinite solution $\bar{Q}$ if and only if:*

(i) *F_1 is asymptotically stable ($|\mu_i(F_1)| < 1$, for all i);*

(ii) *$|\mu_i(F_3)| \neq 1$, for all i.*

By looking at Theorems 1 and 5, it is evident that, whenever a unique positive-definite solution $\bar{Q}$ exists, it is also the unique solution.

In order to avoid ambiguity in the statements of the theorems, it is worth noticing that all the assumptions on F_1 and F_3 must be considered only when the corresponding parts do not vanish. For instance, any assumption on F_3 must be dropped when (F, G) is reachable (i.e., $F_1 = F$).

The term "diagonalizable" in the statement of Theorem 3 means that the eigenvalues of F_3 are simple roots of the minimal polynomial.

Notice that some results above can be seen as the discrete-time counterparts of the theorems presented in [19], concerning the positive (semi-)definite solutions of the continuous-time algebraic Lyapunov equation (CALE).

5. Existence and uniqueness conditions for the solutions of the DPLE. This section is concerned with necessary and sufficient conditions for the existence and uniqueness of periodic solutions of the DPLE (1). The derivation of Theorems 7–12 is straightforward, let us bear in mind the correspondence between the DALE and the DPLE pointed out in § 3 (Propositions 5–8), and the results on the DALE derived in § 4 (Theorems 1–6).

THEOREM 7. *The* DPLE (1) *admits a unique T-periodic solution $\tilde{P}(\cdot)$ if and only if $\bar{\Phi}_\tau$ does not have reciprocal eigenvalues.*

THEOREM 8. *The* DPLE (1) *admits a T-periodic solution $\tilde{P}(\cdot)$ if and only if, for each $\mu \in C$, $y \in C^n$, $z \in C^n$ such that $\bar{\Phi}'_\tau y = \mu y$ and $\bar{\Phi}'_\tau z = \mu^{-1}z$, it results in $y^*\bar{W}_\tau z = 0$.*

THEOREM 9. *The* DPLE (1) *admits a T-periodic solution $\tilde{P}(\cdot)$ that is positive definite for all t if and only if*

(i) *Φ_1 is asymptotically stable ($|\mu_i(\Phi_1)| < 1$, for all i);*

(ii) *$|\mu_i(\Phi_3)| = 1$, for all i;*

(iii) *Φ_3 is diagonalizable;*

(iv) *$X_r(t) = X_c(t)$, for all t.*

THEOREM 10. *The* DPLE (1) *admits a T-periodic solution $\tilde{P}(\cdot)$ that is positive semidefinite for all t if and only if Φ_1 is asymptotically stable ($|\mu_i(\Phi_1)| < 1$, for all i).*

THEOREM 11. *The* DPLE (1) *admits a unique T-periodic positive definite solution $\tilde{P}(\cdot)$ if and only if*

(i) $A(\cdot)$ *is asymptotically stable* ($|\mu_i(\bar{\Phi}_\tau)| < 1$, *for all* i);

(ii) $(A(\cdot), B(\cdot))$ *is reachable at any time point* t.

THEOREM 12. *The* DPLE (1) *admits a unique T-periodic positive semidefinite solution* $\tilde{P}(\cdot)$ *if and only if*

(i) Φ_1 *is asymptotically stable* ($|\mu_i(\Phi_1)| < 1$, *for all* i);

(ii) $|\mu_i(\Phi_3)| \neq 1$, *for all* i.

In view of Theorems 11 and 12, the following results can also be easily obtained.

THEOREM 13. *Suppose that the pair* $(A(\cdot), B(\cdot))$ *is reachable at any time point. Then* $A(\cdot)$ *is asymptotically stable if and only if the* DPLE (1) *admits a unique T-periodic positive-definite solution* $\tilde{P}(\cdot)$.

THEOREM 14. *Suppose that the pair* $(A(\cdot), B(\cdot))$ *is stabilizable. Then,* $A(\cdot)$ *is asymptotically stable if and only if the* DPLE (1) *admits a unique T-periodic positive-semidefinite solution* $\tilde{P}(\cdot)$.

In order to prove Theorem 14, recall that the assumption of stabilizability of $(A(\cdot), B(\cdot))$ implies that all the eigenvalues of Φ_3 lie inside the open unit disk (see, e.g., [16]).

The conditions stated in Theorems 13 and 14, linking the stability of $A(\cdot)$ with the existence of periodic solutions of (1), are usually referred to in the literature under the heading of "periodic Lyapunov lemma." The same results could be alternatively obtained by following the rationale used in [6] and [7] (where the continuous-time case is mainly considered), or by restricting the analysis carried out in [13], for general time-varying linear systems, to periodic systems.

6. Inertia theorems for the DPLE. To complete the overview on the DPLE, this section is devoted to a brief presentation of the so-called "inertia theory." This theory consists of a number of results linking the inertia (i.e., the number of positive, null, and negative eigenvalues) of any symmetric T-periodic solution of the DPLE with the pattern of eigenvalues of the monodromy matrix of $A(\cdot)$. The interested reader is referred to [10] for a complete discussion on this topic. Here, only the major results are reported.

The following short notation will be used. Given a real square matrix M, the symbols $\sigma_c(M)$, $\delta_c(M)$, and $\pi_c(M)$ will represent the number of eigenvalues of M with negative, zero, and positive real part, respectively. The symbols $\sigma_d(M)$, $\delta_d(M)$, and $\pi_d(M)$ will represent the number of eigenvalues of M with modulus less than, equal to, and greater than 1, respectively.

THEOREM 15. *Let the* DPLE (1) *admit a T-periodic solution* $\tilde{P}(\cdot)$.

(i) *If* $(A(\cdot), B(\cdot))$ *is reachable at* t, *then*

$$\pi_c(\tilde{P}(t)) = \sigma_d(\bar{\Phi}_0),$$
$$\sigma_c(\tilde{P}(t)) = \pi_d(\bar{\Phi}_0),$$
$$\delta_c(\tilde{P}(t)) = \delta_d(\bar{\Phi}_0) = 0.$$

(ii) *If* $(A(\cdot), B(\cdot))$ *is controllable, then, for any* t

$$\pi_c(\tilde{P}(t)) = \sigma_d(\bar{\Phi}_0) - q_t,$$
$$\sigma_c(\tilde{P}(t)) = \pi_d(\bar{\Phi}_0),$$
$$\delta_c(\tilde{P}(t)) = q_t,$$
$$\delta_d(\bar{\Phi}_0) = 0$$

where q_t *is the dimension of the unreachability subspace of* $(A(\cdot), B(\cdot))$ *at time* t.

(iii) *If* $(A(\cdot), B(\cdot))$ *is stabilizable, then, for any* t

$$\pi_c(\tilde{P}(t)) + \delta_c(\tilde{P}(t)) = \sigma_d(\bar{\Phi}_0),$$
$$\sigma_c(\tilde{P}(t)) = \pi_d(\bar{\Phi}_0),$$
$$\delta_d(\bar{\Phi}_0) = 0.$$

(iv) *If* $(A(\cdot), B(\cdot))$ *is stabilizable and* $\tilde{P}(\cdot)$ *is the unique T-periodic solution, then, for any* t, *the conclusions of point* (ii) *hold.*

The proof of Theorem 15 in [10] relies on the correspondence between the DPLE (1) and the DALE (4), already mentioned in § 3, and the inertia theory for the DALE.

Notice that some results of Theorem 15, when specialized to the case of positive (semi-)definite solutions, could be obtained as well starting from the necessary conditions of Theorems 9–12 of § 5. For instance, consider Theorem 15(iii). If $\tilde{P}(t) \geqq 0$, for all t, and $(A(\cdot), B(\cdot))$ is stabilizable, the conclusion is drawn that $\sigma_d(\bar{\Phi}_0) = n$, or equivalently that $A(\cdot)$ is asymptotically stable. On the other hand, using the "only if" part of Theorem 10 and recalling that the stabilizability of $(A(\cdot), B(\cdot))$ implies the asymptotic stability of Φ_3, the very same conclusion is easily reached.

7. Results for the CPLE. In this final section, the attention is turned to the continuous-time periodic Lyapunov equation (CPLE) introduced in § 1 (see (2)). Since the analysis of the CPLE mimics the discussion on the DPLE presented in the previous sections, the results will be reviewed very concisely; only the major differences between the two cases will be pointed out.

First, the continuous-time linear periodic system underlying the CPLE (2) is described by

$$\dot{x}(t) = A(t)x(t) + B(t)u(t)$$

where $t \in R$, and $A(\cdot)\colon R \to R^{n\times n}$, $B(\cdot)\colon R \to R^{n\times m}$ are periodic matrices of period $T \in R^+$. A survey on the structural properties of such a class of systems is contained in [14]. Here, only some basic definitions and results are needed. The transition matrix $\Phi(t, \tau)$ associated with $A(\cdot)$ is the solution of the differential equation

$$\frac{d\Phi(t,\tau)}{dt} = A(t)\Phi(t,\tau), \qquad t \geqq \tau$$

with initial condition $\Phi(\tau, \tau) = I_n$. The symbol $\bar{\Phi}_\tau$ will be used again to denote the monodromy matrix of $A(\cdot)$ at τ ($\bar{\Phi}_\tau = \Phi(\tau + T, \tau)$).

The reachability Gramian matrix associated with $(A(\cdot), B(\cdot))$ is

$$W(\tau,t) = \int_\tau^t \Phi(t,\sigma)B(\sigma)B(\sigma)'\Phi(t,\sigma)'d\sigma, \qquad t \geqq \tau$$

and $\bar{W}_\tau = W(\tau, \tau + T)$ represents the single-period reachability Gramian. Furthermore, let $\bar{D}_\tau$ be any matrix such that $\bar{D}_\tau\bar{D}_\tau' = \bar{W}_\tau$. As in the discrete-time case, the periodic system (or, equivalently, $A(\cdot)$) is asymptotically stable if and only if all the eigenvalues of $\bar{\Phi}_\tau$ lie inside the open unit disk. Similarly, the structural properties (reachability, controllability, stabilizability) of $(A(\cdot), B(\cdot))$ can be analyzed in terms of the properties of the discrete-time constant pair $(\bar{\Phi}_\tau, \bar{D}_\tau)$.

Two important results peculiar to the continuous-time case are that: (i) reachability and controllability are equivalent notions; (ii) reachability at a given τ implies reachability at any time point. In conclusion, the results expressed in Propositions 1–4 of § 2 have the following continuous-time counterpart.

PROPOSITION 9. (i) *The continuous-time pair* $(A(\cdot), B(\cdot))$ *is controllable* (*stabilizable*) *if and only if* $(\bar{\Phi}_\tau, \bar{D}_\tau)$ *is controllable* (*stabilizable*).

(ii) $X_r(\tau) = X_c(\tau) = Z_{\tau r} = Z_{\tau c}$.

As for the Kalman canonical decomposition, everything remains formally unchanged with respect to the discrete-time case, with the additional property that the controllable

pair $(A_1(\cdot), B_1(\cdot))$ is also reachable. Keeping the same notation as before, Φ_1 and Φ_3 are the monodromy matrices of the controllable and uncontrollable parts of $(A(\cdot), B(\cdot))$, respectively.

It is easy to verify that the solution at time t of the CPLE (2) starting with $P(\tau) = P_\tau$ is given by

$$P(t) = \Phi(t,\tau)P_\tau\Phi(t,\tau)' + W(\tau,t).$$

When we impose periodicity, it follows that the periodic generator P_τ for the CPLE must satisfy a DALE that is formally identical to (4). Hence, the analysis carried out for the DPLE can be repeated for the CPLE. As a noticeable difference, remark that a T-periodic solution $\tilde{P}(\cdot)$ that is positive definite at a given τ is positive definite at any time point (due to the fact that $\Phi(t, \tau)$ is always nonsingular). The results for the CPLE can be summarized as follows.

THEOREM 16. *The* CPLE (2) *admits a unique T-periodic solution* $\tilde{P}(\cdot)$ *if and only if* $\bar{\Phi}_\tau$ *does not have reciprocal eigenvalues.*

THEOREM 17. *The* CPLE (2) *admits a T-periodic solution* $\tilde{P}(\cdot)$ *if and only if, for each* $\mu \in C$, $y \in C^n$, $z \in C^n$ *such that* $\bar{\Phi}'_\tau y = \mu y$ *and* $\bar{\Phi}'_\tau z = \mu^{-1} z$, *it results in* $y^* \bar{W}_\tau z = 0$.

THEOREM 18. *The* CPLE (2) *admits a T-periodic solution* $\tilde{P}(\cdot)$ *that is positive definite for all t if and only if*

(i) Φ_1 *is asymptotically stable* ($|\mu_i(\Phi_1)| < 1$, *for all* i);

(ii) $|\mu_i(\Phi_3)| = 1$, *for all* i;

(iii) Φ_3 *is diagonalizable.*

THEOREM 19. *The* CPLE (2) *admits a T-periodic solution* $\tilde{P}(\cdot)$ *that is positive semidefinite for all t if and only if* Φ_1 *is asymptotically stable* ($|\mu_i(\Phi_1)| < 1$, *for all* i).

THEOREM 20. *The* CPLE (2) *admits a unique T-periodic positive-definite solution* $\tilde{P}(\cdot)$ *if and only if*

(i) $A(\cdot)$ *is asymptotically stable* ($|\mu_i(\bar{\Phi}_\tau)| < 1$, *for all* i);

(ii) $(A(\cdot), B(\cdot))$ *is controllable.*

THEOREM 21. *The* CPLE (2) *admits a unique T-periodic positive semidefinite solution* $\tilde{P}(\cdot)$ *if and only if*:

(i) Φ_1 *is asymptotically stable* ($|\mu_i(\Phi_1)| < 1$, *for all* i);

(ii) $|\mu_i(\Phi_3)| \neq 1$, *for all* i.

THEOREM 22 (Lyapunov Lemma). *Suppose that the continuous-time pair* $(A(\cdot), B(\cdot))$ *is controllable. Then* $A(\cdot)$ *is asymptotically stable if and only if the* CPLE (2) *admits a unique T-periodic positive definite solution* $\tilde{P}(\cdot)$.

THEOREM 23 (Extended Lyapunov Lemma). *Suppose that the continuous-time pair* $(A(\cdot), B(\cdot))$ *is stabilizable. Then,* $A(\cdot)$ *is asymptotically stable if and only if the* CPLE (2) *admits a unique T-periodic positive semidefinite solution* $\tilde{P}(\cdot)$.

THEOREM 24 (Inertia theorem). *Let the* CPLE (2) *admit a T-periodic solution* $\tilde{P}(\cdot)$.

(i) *If* $(A(\cdot), B(\cdot))$ *is controllable, then for any t*
$\pi_c(\tilde{P}(t)) = \sigma_d(\bar{\Phi}_0)$,
$\sigma_c(\tilde{P}(t)) = \pi_d(\bar{\Phi}_0)$,
$\delta_c(\tilde{P}(t)) = \delta_d(\bar{\Phi}_0) = 0$.

(ii) *If* $(A(\cdot), B(\cdot))$ *is stabilizable, then for any t*
$\pi_c(\tilde{P}(t)) + \delta_c(\tilde{P}(t)) = \sigma_d(\bar{\Phi}_0)$,
$\sigma_c(\tilde{P}(t)) = \pi_d(\bar{\Phi}_0)$,
$\delta_d(\bar{\Phi}_0) = 0$.

(iii) *If* $(A(\cdot), B(\cdot))$ *is stabilizable and* $\tilde{P}(\cdot)$ *is the unique T-periodic solution, then for any t*

$\pi_c(\tilde{P}(t)) = \sigma_d(\bar{\Phi}_0) - q,$

$\sigma_c(\tilde{P}(t)) = \pi_d(\bar{\Phi}_0),$

$\delta_c(\tilde{P}(t)) = q,$

$\delta_d(\bar{\Phi}_0) = 0$

where q is the dimension of the uncontrollability subspace of $(A(\cdot), B(\cdot))$. □

For an alternative derivation of Theorems 22 and 23, see [7]. The proof of Theorem 24 can be found in [8] and [9] (see also [11]).

Appendix. This Appendix contains the proofs of the results presented in § 4.

The proofs of Theorems 1 and 2 are based on the Kronecker calculus (see, e.g., [20]). Precisely, the DALE (5) can be rewritten in the standard form:

$$Hq = g \tag{8}$$

where $q = \mathrm{Vec}(Q)$, $g = \mathrm{Vec}(GG')$, and $H = I_{n^2} - F \otimes F$. The existence and uniqueness conditions for the solutions $\bar{q}$ of (8) are then reinterpreted in terms of F and G.

Proof of Theorem 1. The eigenvalues of H are given by $1 - \mu_i(F)\mu_j(F)$ for all the pairs i, j. Hence, (8) admits a unique solution $\bar{q}$ (H is nonsingular) if and only if $\mu_i(F)\mu_j(F) \neq 1$, for all i, j. □

Proof of Theorem 2. A necessary and sufficient condition for the existence of a solution q of (8) is that $g \in R[H] = N[H']^{\perp}$. This is equivalent to saying that for all $v \in R^{n^2}$, $v'H = 0$ implies $v'g = 0$. It is known (see [20, p. 27]) that any vector v satisfying $v'H = 0$ is given by $v = z \otimes y$, where $y \in C^n$ and $z \in C^n$ are such that $F'y = \mu y$ and $F'z = \mu^{-1}z$ for some $\mu \in C$. Then, by using standard properties of the Kronecker product, the following results:

$$v'g = (z \otimes y)'\mathrm{Vec}(GG') = (z^* \otimes y^*)(G \otimes G)\mathrm{Vec}(I_n)$$

$$= (z^*G) \otimes (y^*G)\mathrm{Vec}(I_n) = \mathrm{Vec}(y^*GG'z) = y^*GG'z.$$

The result of Theorem 2 easily follows. □

In the derivation of Theorems 3 and 4, reference will be made to the canonical decomposition (7). According to such a decomposition, the DALE (5) can be split into the following subequations:

$$Q_1 = F_1Q_1F_1' + F_2Q_2'F_1' + F_1Q_2F_2' + F_2Q_3F_2' + G_1G_1', \tag{9}$$

$$Q_2 = F_1Q_2F_3' + F_2Q_3F_3', \tag{10}$$

$$Q_3 = F_3Q_3F_3', \tag{11}$$

with

$$Q = \begin{bmatrix} Q_1 & Q_2 \\ Q_2' & Q_3 \end{bmatrix}.$$

Proof of Theorem 3. (Necessity.) Let $\bar{Q} > 0$ be a solution of the DALE (5). Now consider an eigenvalue μ of F, and let $z \in C^n$ be an associate eigenvector of F' ($F'z = \mu z$). By premultiplying both sides of (5) by z^* and postmultiplying them by z, we obtain

$$(|\mu|^2 - 1)z^*\bar{Q}z + z^*GG'z = 0.$$

Since $z^*\bar{Q}z > 0$, it results in $|\mu| \leqq 1$. Hence, F does not admit eigenvalues outside the unit disk. Due to this fact, we can equivalently consider the following partitioned form of F:

$$\tilde{F} = \begin{bmatrix} F_a & 0 \\ 0 & F_b \end{bmatrix}$$

where all the eigenvalues of F_a lie inside the open unit disk and all the eigenvalues of F_b lie on the unit circle. The corresponding matrix $\tilde{G}$ obtained from G is

$$\tilde{G} = \begin{bmatrix} G_a \\ G_b \end{bmatrix}$$

and the solution $\bar{Q} > 0$ is transformed into

$$\tilde{Q} = \begin{bmatrix} Q_a & Q_c \\ Q_c' & Q_b \end{bmatrix} > 0.$$

Equation (5) can be split into

$$Q_a = F_a Q_a F_a' + G_a G_a', \tag{12}$$

$$Q_c = F_a Q_c F_b' + G_a G_b',$$

$$Q_b = F_b Q_b F_b' + G_b G_b'. \tag{13}$$

We now show that F_b is diagonalizable. Indeed, let y be a complex vector such that $F_b' y = \mu y$, $\mu \in C$. Recalling that $|\mu| = 1$, (13) implies that

$$(|\mu|^2 - 1) y^* Q_b y + y^* G_b G_b' y = y^* G_b G_b' y = 0.$$

Hence, $G_b' y = 0$. Suppose now by contradiction that F_b is not diagonalizable. Then, for at least one eigenvalue μ of F_b, there exists a nonzero generalized eigenvector z such that $F_b' y = \mu y$ and $F_b' z = \mu z + y$. From (13) and $G_b' y = 0$, it follows that

$$\begin{aligned} 0 &= z^* F_b Q_b F_b' y - z^* Q_b y + z^* G_b G_b' y \\ &= (\mu^* z^* + y^*) Q_b \mu y - z^* Q_b y \\ &= (|\mu|^2 - 1) z^* Q_b y + \mu y^* Q_b y = \mu y^* Q_b y. \end{aligned}$$

This contradicts $Q_b > 0$ and, in turn, the assumption $\bar{Q} > 0$. Consequently, F_b is diagonalizable and its eigenvectors y_i form a basis of C^{n_b} (n_b being the order of F_b). Hence, $G_b' y_i = 0$, for all i implies $G_b = 0$.

It is then apparent that the subspace Y spanned by the vectors $[0\ x']'$, $x \in R^{n_b}$ is contained in the unreachability subspace $\tilde{Z}_{ur}$ of the pair $(\tilde{F}, \tilde{G})$. Since F_b does not have null eigenvalues, Y is also contained in the uncontrollability subspace $\tilde{Z}_{uc}$ (see, e.g., [21]). In particular, $n_b \leqq \dim(\tilde{Z}_{uc})$.

Now, consider the DALE (5) decomposed as in (9)–(11). Since there exists a solution $\bar{Q}_3 > 0$, it is easy to show that all the eigenvalues of the uncontrollable part (F_3) lie on the unit circle. Indeed, letting μ be an eigenvalue of F_3 and z be an associate eigenvector of F_3', (11) implies

$$(|\mu|^2 - 1) z^* \bar{Q}_3 z = 0.$$

Since $z^* \bar{Q}_3 z > 0$, this equation is verified only if $|\mu| = 1$.

Therefore, the uncontrollability subspace $\hat{Z}_{uc}$ of $(\hat{F}, \hat{G})$ is such that dim $(\hat{Z}_{uc}) \leqq n_b$. Since dim $(\tilde{Z}_{uc})$ = dim $(\hat{Z}_{uc})$, the conclusion is drawn that n_b = dim $(\tilde{Z}_{uc})$ and $Y \equiv \tilde{Z}_{uc}$. Thus, the pair $(\tilde{F}, \tilde{G})$ is a particular canonical decomposition. Hence, F_a is similar to F_1 and F_b is similar to F_3. This completes the proof of points (i)–(iii).

As for point (iv), it suffices to prove that the pair (F_a, G_a), besides being controllable, is also reachable (in view of the discussion above, (F_1, G_1) and (F_a, G_a) are equivalent pairs). Suppose by contradiction that (F_a, G_a) is not reachable. Then, bearing in mind the well-known reachability and controllability PBH tests (see, e.g., [15]), we find that there exists a complex vector x such that $F_a'x = 0$ and $G_a'x = 0$. By (12), this leads to $x^*Q_a x = 0$, which violates the hypothesis $\tilde{Q} > 0$. □

Proof of Theorem 3. (Sufficiency.) From assumptions (i) and (ii), the matrices F_1 and F_3 do not have common eigenvalues. Thus there exists a nonsingular $n \times n$ transformation

$$S = \begin{bmatrix} I & \bar{S} \\ 0 & I \end{bmatrix}$$

(where $\bar{S}$ has the same dimensions as F_2) such that

$$\tilde{F} = S\hat{F}S^{-1} = \begin{bmatrix} F_1 & 0 \\ 0 & F_3 \end{bmatrix}, \qquad \tilde{G} = S\hat{G} = \begin{bmatrix} G_1 \\ 0 \end{bmatrix}$$

with $F_2 + \bar{S}F_3 = F_1\bar{S}$. As a matter of fact, this last equation admits a (unique) solution $\bar{S}$ due to the fact that F_1 and F_3 do not have common eigenvalues (see, e.g., [20]).

When the DALE (5) is reformulated in terms of $(\tilde{F}, \tilde{G})$, any solution

$$\tilde{Q} = \begin{bmatrix} Q_1 & Q_2 \\ Q_2' & Q_3 \end{bmatrix}$$

satisfies

$$\begin{aligned} Q_1 &= F_1 Q_1 F_1' + G_1 G_1', \\ Q_2 &= F_1 Q_2 F_3', \\ Q_3 &= F_3 Q_3 F_3'. \end{aligned} \tag{14}$$

It is a simple matter of matrix algebra to recognize that F_3 is similar to an orthogonal matrix, namely, that there exists a square matrix $\hat{S}$ such that $\hat{S}^{-1}F_3\hat{S}$ is orthogonal. Hence, a solution $\tilde{Q} > 0$ can be constructed by taking $Q_3 = \alpha\hat{S}\hat{S}^*$, $\alpha > 0$, $Q_2 = 0$, and $Q_1 > 0$ satisfying (14). From assumption (iv), the pair (F_1, G_1) is reachable. Moreover, F_1 is asymptotically stable. Hence, (14) admits a positive definite solution Q_1 in view of the Lyapunov lemma (see [22]). □

Proof of Theorem 4. (Necessity.) Let $\bar{Q} \geqq 0$ be a solution of the DALE (5). It is well known that, given $\bar{Q} \geqq 0$, there exists a nonsingular transformation S such that

$$\check{Q} = S\bar{Q}S' = \begin{bmatrix} Q_r & 0 \\ 0 & 0 \end{bmatrix}$$

with $Q_r > 0$. It is easy to see that $\check{Q}$ is a solution of

$$Q = \check{F}Q\check{F}' + \check{G}\check{G}'$$

where $\check{F}$ and $\check{G}$ are defined as

$$\check{F} = SFS^{-1} = \begin{bmatrix} F_r & F_s \\ F_t & F_u \end{bmatrix}, \qquad \check{G} = SG = \begin{bmatrix} G_r \\ G_t \end{bmatrix}.$$

It can be shown by direct computation that

$$Q_r = F_r Q_r F_r' + G_r G_r', \tag{15}$$

$$0 = F_t Q_r F_t' + G_t G_t'. \tag{16}$$

Since $Q_r > 0$, from (16) it is apparent that $F_t = 0$ and $G_t = 0$. Hence, the subspace Y spanned by the vectors $[0\ x']'$, $x \in R^{n_u}$ (where n_u is the order of F_u), is contained in the unreachability subspace $\check{Z}_{ur}$ of the pair $(\check{F}, \check{G})$. A moment's reflection reveals that the eigenvalues of the reachable part of $(\check{F}, \check{G})$ coincide with those of the reachable part of (F_r, G_r).

Now consider (15). In view of Theorem 3, since (15) admits a solution $Q_r > 0$, the eigenvalues of the reachable part of (F_r, G_r) lie inside the unit circle. Thus, the reachable part of $(\check{F}, \check{G})$ is asymptotically stable as well. Recall that the asymptotic stability of the reachable part is equivalent to that of the controllable part (see, e.g., [21]); thus the proof is completed. □

Proof of Theorem 4 (*Sufficiency*). Consider the decomposition (9)–(11) of the DALE (5). It is apparent that $\bar{Q}_3 = 0$ and $\bar{Q}_2 = 0$ solve (10) and (11). Hence, (9) reduces to the DALE

$$Q_1 = F_1 Q_1 F_1' + G_1 G_1'$$

with (F_1, G_1) controllable. Since F_1 is asymptotically stable by hypothesis, the existence of a solution $\bar{Q}_1 \geqq 0$ is ensured by the Lyapunov Lemma (see, e.g., [22]). In conclusion, the overall solution $\bar{Q}$ defined by $\bar{Q}_1$, $\bar{Q}_2$, and $\bar{Q}_3$ is positive semidefinite. □

Proof of Theorem 5. (Necessity.) Suppose that there exists a unique solution $\bar{Q} > 0$ of the DALE (5). Then, in particular, Theorem 3 yields $|\mu_i(F_1)| < 1$, for all i, $|\mu_i(F_3)| = 1$, for all i, F_3 is diagonalizable, and (F_1, G_1) is reachable. Consider again the transformation used in the sufficient part of Theorem 3. It was shown there that a positive definite solution of the transformed DALE is given by

$$\tilde{Q} = \begin{bmatrix} Q_1 & 0 \\ 0 & \alpha \hat{S} \hat{S}^* \end{bmatrix}$$

where α is an arbitrary positive constant, $\hat{S}$ is such that $\hat{S}^{-1} F_3 \hat{S}$ is orthogonal, and $Q_1 > 0$ satisfies (14). The uniqueness assumption obviously implies that F_3 must vanish. Thus, $F = F_1$, and the proof is complete. □

Proof of Theorem 5. (Sufficiency.) The proof is part of the well-known Lyapunov Lemma [22]. □

Proof of Theorem 6. (Necessity.) Suppose that there exists a unique positive semidefinite solution $\bar{Q}$ of the DALE (5). In view of Theorem 4, the controllable part (F_1) is asymptotically stable, so that point (i) is established. By a suitable transformation, the pair (F, G) can be put in the following form:

$$\hat{\check{F}} = \begin{bmatrix} F_c & 0 & 0 \\ 0 & F_d & 0 \\ 0 & 0 & F_e \end{bmatrix}, \qquad \hat{\check{G}} = \begin{bmatrix} G_c \\ 0 \\ 0 \end{bmatrix}$$

where the eigenvalues of F_c, F_d, and F_e lie inside the unit disk, on the unit circle, and

outside the unit disk, respectively. The existence of zero elements in $\hat{\hat{G}}$ can be easily inferred from the asymptotic stability of the controllable part. In consideration of the transformed DALE, any solution

$$\hat{\hat{Q}} = \begin{bmatrix} Q_c & Q_f & Q_g \\ Q_f' & Q_d & Q_h \\ Q_g' & Q_h' & Q_e \end{bmatrix}$$

satisfies

$$Q_c = F_c Q_c F_c' + G_c G_c', \tag{17}$$

$$Q_f = F_c Q_f F_d', \tag{18}$$

$$Q_d = F_d Q_d F_d', \tag{19}$$

$$Q_e = F_e Q_e F_e'. \tag{20}$$

Since $|\mu_i(F_e)| > 1$, for all i, the unique solution of (20) is $Q_e = 0$. Since $\mu_i(F_c) \neq \mu_j(F_d)^{-1}$, for all i, j, the unique solution of (18) is $Q_f = 0$. Thus, any positive semidefinite solution must take on the form

$$\hat{\hat{Q}} = \begin{bmatrix} Q_c & 0 & 0 \\ 0 & Q_d & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

with $Q_c \geqq 0$ and $Q_d \geqq 0$ satisfying (17) and (19), respectively. It is easy to see that, in view of the fact that $|\mu_i(F_d)| = 1$ for all i, (19) admits an infinite number of positive semidefinite solutions (the Jordan form of F_d can be considered to draw this conclusion). Hence, in order to result in uniqueness of the overall solution, F_d must vanish. This leads to the result that $|\mu_i(F)| \neq 1$, for all i, and, in particular, $|\mu_i(F_3)| \neq 1$, for all i (which proves point (ii)). □

Proof of Theorem 6. (Sufficiency.) Assume that conditions (i) and (ii) are verified. From Theorem 4, the DALE (5) admits a solution $\bar{Q} \geqq 0$. Moreover, since $|\mu_i(F)| \neq 1$, for all i, we can consider the following decomposition of F and G:

$$\check{\check{F}} = \begin{bmatrix} F_v & 0 \\ 0 & F_z \end{bmatrix}, \qquad \check{\check{G}} = \begin{bmatrix} G_v \\ 0 \end{bmatrix}$$

where the eigenvalues of F_v and F_z lie inside and outside the unit disk, respectively. Any solution

$$\check{\check{Q}} = \begin{bmatrix} Q_v & Q_w \\ Q_w' & Q_z \end{bmatrix}$$

of the transformed DALE satisfies

$$Q_v = F_v Q_v F_v' + G_v G_v', \tag{21}$$

$$Q_w = F_v Q_w F_z', \tag{22}$$

$$Q_z = F_z Q_z F_z'. \tag{23}$$

The unique solution of (23) is $Q_z = 0$. Moreover, in order to have $\check{\check{Q}} \geqq 0$, we must have $Q_w = 0$ also. Finally, (21) admits a unique solution Q_v, which is positive semidefinite. Hence, the DALE (5) admits a unique positive semidefinite solution. □

REFERENCES

[1] M. A. LYAPUNOV, *Problème general de la stabilité du mouvement*, Ann. Fac. Sci. Toulouse Math. (5), 9 (1907), pp. 203–474.

[2] R. E. KALMAN AND J. E. BERTRAM, *Control system analysis and design by the second method of Lyapunov*, Trans. ASME Ser. D J. Basic Engrg., 82 (1960), pp. 371–400.

[3] R. E. KALMAN, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Nat. Acad. Sci. U.S.A., 49 (1963), pp. 201–205.

[4] A. OSTROWSKI AND H. SCHNEIDER, *Some theorems on the inertia of general matrices*, J. Math. Anal. Appl., 4 (1962), pp. 72–84.

[5] H. K. WIMMER AND A. D. ZIEBUR, *Remarks on inertia theorems for matrices*, Czechoslovak Math. J., 25 (1975), pp. 556–561.

[6] S. BITTANTI, P. BOLZERN, AND P. COLANERI, *Stability analysis of linear periodic systems via the Lyapunov equation*, in Proc. 9th IFAC World Congress, Budapest, Hungary, pp. 169–172, 1984.

[7] ———, *The extended periodic Lyapunov lemma*, Automatica, 21 (1985), pp. 603–605.

[8] M. A. SHAYMAN, *Inertia theorems for the periodic Lyapunov equation and periodic Riccati equation*, Systems Control Lett., 4 (1984), pp. 27–32.

[9] S. BITTANTI AND P. COLANERI, *Lyapunov and Riccati equations: periodic inertia theorems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 659–661.

[10] P. BOLZERN AND P. COLANERI, *Inertia theorems for the periodic Lyapunov difference equation and the periodic Riccati difference equation*, Linear Algebra Appl., 85 (1987), pp. 249–265.

[11] S. BITTANTI, P. BOLZERN, AND P. COLANERI, *Inertia theorems for Lyapunov and Riccati equations—an updated view*, presented at SIAM Conference on Linear Algebra in Signals, Systems, and Control, Boston, MA, 1986.

[12] B. D. O. ANDERSON AND J. B. MOORE, *New results in linear systems stability*, SIAM J. Control, 7 (1969), pp. 398–414.

[13] ———, *Detectability and stabilizability of time-varying discrete-time linear systems*, SIAM J. Control Optim., 19 (1981), pp. 20–32.

[14] S. BITTANTI, *Deterministic and stochastic linear periodic systems*, in Time Series and Linear Systems, S. Bittanti ed., Springer-Verlag, Berlin, New York, 1986, pp. 141–182.

[15] T. KAILATH, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[16] P. BOLZERN, *Criteria for reachability, controllability and stabilizability of discrete-time linear periodic systems*, in Proc. 5th Polish-English Seminar on Real-Time Process Control, Radziejowice, Poland, 1986, pp. 69–83.

[17] S. BITTANTI, P. COLANERI, AND G. DE NICOLAO, *Discrete-time linear periodic systems: a note on the reachability and controllability interval length*, Systems Control Lett., 8 (1986), pp. 75–78.

[18] R. A. MEYER AND C. S. BURRUS, *A unified analysis of multirate and periodically time-varying digital filters*, IEEE Trans. Circuits and Systems, CAS-22 (1975), pp. 162–168.

[19] J. SNYDERS AND M. ZAKAI, *On nonnegative solutions of the equation $AD + DA' = -C$*, SIAM J. Appl. Math., 18 (1970), pp. 704–714.

[20] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, Ellis Horwood, Chichester, 1981.

[21] O. M. GRASSELLI, *Conditions for controllability and reconstructibility of discrete-time linear composite systems*, Internat. J. Control, 31 (1980), pp. 433–441.

[22] B. D. O. ANDERSON AND J. B. MOORE, *Optimal Filtering*, Prentice–Hall, Englewood Cliffs, NJ, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 4, October 1988

BLOCK-SEQUENTIAL ALGORITHMS FOR SET-THEORETIC ESTIMATION*

RONALD K. PEARSON†

Abstract. A new algorithm is proposed for solving the set-theoretic parameter estimation problem. In contrast to point estimation strategies like least squares or maximum likelihood, the set-theoretic parameter estimation problem imposes bounds on model errors and seeks the resulting bounds imposed on the free model parameters. The exact solution to this problem is a convex polytope in the parameter space with too many vertices for an exact solution to be practical. Thus, the standard solution approach is to seek an outer bounding set that is more easily parameterized. This paper describes a computational approach that interpolates in estimation efficiency and computational effort between two extreme cases described by other authors. Two simple numerical examples are included.

Key words. set-theoretic estimation, unknown-but-bounded uncertainty, parameter estimation algorithms, simultaneous linear inequalities, confidence regions

AMS(MOS) subject classifications. 52A40, 62F25, 93B30

1. Introduction.

Given a sequence of N experimentally measured "dependent variables" $\{y_i\}$ and an assumed linear model of the form

$$y_i = \sum_{j=1}^{p} G_{ij} a_j + e_i, \qquad i = 1, 2, \cdots, N, \tag{1.1}$$

the linear parameter estimation problem is to determine values for the unknown model parameters $\{a_j\}$. Here, $\{G_{ij}\}$ is a set of $N \times p$ "independent variables" defined by the specific model structure chosen (cf. § 5 for a more detailed discussion) and $\{e_i\}$ is the sequence of modeling errors that represents the degree of mismatch between the available data and the assumed model. Deterministic point estimation strategies (e.g., least squares) proceed by minimizing some norm of $\{e_i\}$, yielding "optimal" estimates of the unknown parameters $\{a_j\}$, but providing no information about parameter uncertainty. If a statistical description of the modeling errors is assumed, confidence intervals may be computed, but this approach is not feasible when the statistics are unknown or when statistical descriptions are inappropriate [1].

In such cases, a weaker model error assumption that is often appropriate is the set-theoretic or "unknown but bounded" [2] error constraint

$$e_i^- \leqq e_i \leqq e_i^+, \qquad i = 1, 2, \cdots, N \tag{1.2}$$

where the bounds $\{e_i^-\}$ and $\{e_i^+\}$ are known. Given these bounds, the set-theoretic parameter estimation problem is to determine the region S^* in parameter space R^p containing the parameter vectors $\mathbf{a} = [a_1, \cdots, a_p]^T$ consistent with (1.1) and (1.2).

2. Exact and approximate estimate sets.

To proceed, note that the model (1.1) and the error condition (1.2) may be combined to yield the $2N$ simultaneous inequalities

$$y_i - e_i^+ \leqq \sum_{j=1}^{p} G_{ij} a_j \leqq y_i - e_i^-, \qquad i = 1, 2, \cdots, N, \tag{2.1}$$

* Received by the editors August 25, 1986; accepted for publication (in revised form) February 29, 1988.
† Engineering Physics Laboratory, Experimental Station, E.I. duPont de Nemours & Co., Inc., Wilmington, Delaware 19898.

that must be satisfied by all parameter vectors $\mathbf{a}$ in the admissible parameter set S^*. Condition (2.1) implies that S^* is the intersection of $2N$ closed halfspaces in R^p, which defines a polyhedral set [3, p. 26]. Further, if S^* is bounded, it is a convex polytope in R^p [3, p. 32], defined as the convex hull of a finite number of points. In practice, however, this number is much too large for an exact determination of S^* to be feasible, so some computationally simpler approximation strategy is necessary.

The most common approximation strategy is to seek an outer bounding set $O^* \supseteq S^*$ that belongs to some class $C(c_0, \Sigma)$ of sets in R^p parameterized by a p-vector c_0 and a nonsingular $p \times p$ matrix Σ. The class C defines the shape of the bounding sets considered, the vector c_0 defines their center positions, and the matrix Σ defines their size and orientation. Four specific classes of this type are the following:

(1) Ellipsoids in R^p:

$$E(c_0, \Sigma) = \{x \in R^p \mid (x - c_0)^T \Sigma (x - c_0) \leqq 1, \Sigma = \Sigma^T, \Sigma > 0\};$$

(2) Parallelepipeds in R^p:

$$P(c_0, \Sigma) = \{x \in R^p \mid \|\Sigma(x - c_0)\|_\infty \leqq 1\};$$

(3) Rectangular parallelepipeds in R^p:

$$R(c_0, \Sigma) = \{x \in R^p \mid \|\Sigma(x - c_0)\|_\infty \leqq 1, \text{rows of } \Sigma \text{ orthogonal}\};$$

(4) Rectangular parallelepipeds in R^p, oriented parallel to coordinate axes:

$$PR(c_0, \Sigma) = \{x \in R^p \mid \|\Sigma(x - c_0)\|_\infty \leqq 1, \Sigma \text{ diagonal}\}.$$

Given a collection $\{O_k\}$ of sets from a class C such that $O_k \supseteq S^*$ for all k, define the volume ratios

$$\omega_k = \text{vol}\{O_k\} / \text{vol}\{S^*\}. \tag{2.2}$$

The bounding set O_k will be termed *tighter* than O_j if $\omega_k < \omega_j$. Further, given a class C of bounding sets, if the quantity

$$\omega_k = \omega^* = \min_{\substack{O \in C \\ O \supseteq S^*}} \text{vol}\{O\} / \text{vol}\{S^*\} \tag{2.3}$$

exists, any bounding set O_k for which $\omega_k = \omega^*$ will be termed *maximally tight*.

Several points should be made regarding this notion of tightness. First, note that $\omega_k \geqq 1$ for any bounding set $O_k \supseteq S^*$. Further, since S^* is a closed, convex set, $\omega_k = 1$ implies $O_k = S^*$. The notion of maximally tight bounding sets defined here is similar to the notion of tightness defined by Kahan [4] for the class $E(c_0, \Sigma)$. Specifically, Kahan considers the intersection I of two ellipsoids and defines a third ellipsoid O to be a tight bound for this intersection if $O \supseteq E \supseteq I$ implies $E = O$ for any fourth ellipsoid E. This notion of tightness may be extended to any class C of bounding sets, i.e., O is tight on S if $O \supseteq X \supseteq S$ implies $X = O$. The following theorem shows that this is a weaker condition than maximal tightness.

THEOREM 2.1. *If O^* is a maximally tight member of the class C of closed, convex bounding sets, then it is tight in the sense of Kahan.*

Proof. Suppose O^* is maximally tight and let X be another member of the class C such that $O^* \supseteq X \supseteq S^*$. Since $O^* \supseteq X$, $\omega_X \leqq \omega^* \Rightarrow \omega_X = \omega^*$. Thus, vol $\{X\}$ = vol $\{O^*\}$ and $O^* \supseteq X$, implying $X = O^*$ since both sets are closed and convex. □

In terms of these notions, the set-theoretic problem considered here consists of three steps:

(i) Select a class $C(\mathbf{c}_0, \Sigma)$ of bounding sets.

(ii) Compute $(\mathbf{c}_0^{\#}, \Sigma^{\#})$ corresponding to the tightest possible bounding set $O^{\#}$ in class C for S^*.

(iii) Relate $(\mathbf{c}_0^{\#}, \Sigma^{\#})$ to the original parameters $\{a_j\}$.

The class $E(c_0, \Sigma)$ has been most commonly used in set-theoretic estimation algorithms because ellipsoids have sufficient geometric flexibility to provide reasonably tight bounds for S^* in step (ii). Ellipsoidal bounding sets suffer from two distinct disadvantages, however. First, the interpretation of the parameter bounds required in step (iii) is complicated by the fact that the defining condition for an ellipsoid consists of p-coupled quadratic inequalities in the parameters $\{a_j\}$. For $p > 3$, direct visualization of this set is impossible, so interpretation of the results can become a significant practical difficulty. The second primary difficulty with the class of ellipsoidal bounding sets is that it is not closed under intersection, introducing complications in sequential set-theoretic estimation procedures in which a priori bounds are to be combined with updated bounds obtained from new data.

The class $P(c_0, \Sigma)$ of bounding parallelepipeds in R^p is similar in its geometric flexibility to the class $E(c_0, \Sigma)$, exhibiting similar advantages and disadvantages, although it does not appear to have been considered before. The classes $R(c_0, \Sigma)$ and $PR(c_0, \Sigma)$ are subsets of $P(c_0, \Sigma)$ obtained by restricting the form of the matrix Σ. Consequently, tighter bounds can generally be obtained in $P(c_0, \Sigma)$ than in $R(c_0, \Sigma)$, while these bounds are correspondingly tighter than those achievable in $PR(c_0, \Sigma)$. Regardless, the class $PR(c_0, \Sigma)$ has been considered by other authors [1], [5], [6] because it is the only one of these classes of bounding sets that does not suffer from the two difficulties noted for ellipsoids. In particular, interpretation of the bounding sets is immediate, since each set represents independent bounds of the form

$$c_0^j - \Sigma_{jj}^{-1} \leqq a_j \leqq c_0^j + \Sigma_{jj}^{-1}. \tag{2.4}$$

Similarly, $PR(c_0, \Sigma)$ is closed under intersection (i.e., $(PR, \cap)$ is a meet semilattice [7]), considerably simplifying the computational effort required to sequentially combine a priori parameter estimates with updated parameter estimates. Consequently, attention will be focused for the remainder of this paper on the class $PR(c_0, \Sigma)$, hereafter called "parallelepiped bounding sets" for simplicity.

3. Rectangular parallelepiped bounding algorithm. Various approaches have been proposed for computing parallelepiped bounding sets [1], [5], [6]. Maximally tight bounding sets are obtained with the algorithm of Milanese and Belforte [1], who reduced the problem to that of solving $2p$ linear programs, each in p variables with $2N$ constraints. The principal disadvantages of this algorithm are the computational effort required for large N and the fact that it is a batch algorithm. Consequently, this algorithm is not suited to real-time applications or exploratory data analysis applications where outlier detection is a significant concern, or to the analysis of nonstationary data where adaptive parameter estimators may be desirable. As a partial solution to these problems, Belforte, Bona, and Cerone [5] use an ellipsoidal algorithm developed by Fogel and Huang [8] to preprocess the data sequentially. In so doing, they obtain a bounding ellipsoid E that intersects all of the hyperplanes defining S^* (i.e., all of the "active constraints"), and possibly some, but not all, of the other hyperplanes defined in (2.1). Thus, by selecting only those hyperplanes that intersect E, they are able to reduce the number of constraints in the Milanese and Belforte algorithm from $2N$ to something substantially smaller.

An even simpler algorithm for computing bounding parallelepipeds has been proposed by Fogel and Huang [6] (this should not be confused with their ellipsoidal algorithm [8]). In this parallelepiped algorithm, the data is processed one point at a time, using

each inequality in (2.1) along with a priori parameter uncertainty intervals to solve for a posteriori parameter uncertainty intervals. Unfortunately, the quality of the parameter estimates obtained with this algorithm is very strongly dependent on the quality of the a priori parameter estimates used to initialize it. This point will be illustrated in the numerical example described in § 6.1.

The problem considered throughout the rest of this paper is that of preliminary or exploratory data analysis, since that is where set-theoretic parameter estimation seems most appropriate. In such cases, the number of model parameters p involved is typically fairly small (say 2 to 10), while the number of data points available may be quite large (e.g., hundreds or thousands). Further, a priori parameter estimates, if available at all, are generally very conservative, and the available data may contain "bad data points" or "outliers." Thus, desirable features in a set-theoretic estimation algorithm for such applications are the following: First, the algorithm should process data sequentially to facilitate detection of outliers, without requiring massive recomputation efforts to correct for their effects if they are present. Second, because N is large, computational complexity should not depend strongly on N, ideally growing only as $O(N)$. Finally, because a priori estimates are generally conservative bounds (and possibly inaccurate ones), the final estimates should not depend on them too strongly.

The algorithm proposed here for computing an outer bounding set containing S^* is a "divide and conquer" strategy that interpolates between the computational efficiency of Fogel and Huang's algorithm and the maximally tight bounds obtained by Milanese and Belforte. Specifically, suppose $N = Kp$ for some integer K, and partition the available data $\{y_i\}$ and $\{G_{ij}\}$ into K disjoint subsets, of size p and $p \times p$, respectively. Alternative partitionings will be considered in § 5, but there is no loss of generality in assuming this partitioning first for simplicity. Inequalities (2.1) then decouple into K sets of inequalities, each of the form

$$u_i \leqq \sum_{j=1}^{p} X_{ij} a_j \leqq v_i, \qquad i = 1, 2, \cdots, p \tag{3.1}$$

where

$$u_i = y_{(k-1)p+i} - e^{+}_{(k-1)p+i}, \tag{3.2a}$$

$$v_i = y_{(k-1)p+i} - e^{-}_{(k-1)p+i}, \tag{3.2b}$$

$$X_{ij} = G_{(k-1)p+i,j} \tag{3.2c}$$

for $i = 1, 2, \cdots, p$, $j = 1, 2, \cdots, p$, and $k = 1, 2, \cdots, K$. Each of these K sets of $p \times p$ simultaneous inequalities defines bounding sets S_k whose intersection yields the exact solution set S^*. Consequently, if outer bounding sets O_k are computed for each solution set S_k, their intersection yields an overall bounding set O^* for the complete solution set S^*. In general, this bounding set will be looser than that obtained by Milanese and Belforte's algorithm, but tighter than that obtained by Fogel and Huang's parallelepiped algorithm.

The global computational strategy just described reduces the problem of computing the bounding set O^* to one of solving K sets of $p \times p$ simultaneous linear inequalities. A variety of computational strategies may be applied to these, an obvious possibility being the algorithm of Milanese and Belforte with $N = p$, obtaining the solution by linear programming. The approach pursued here, however, is to exploit the structure of (3.1) to obtain a more efficient algorithm. One possibility would be to develop a variation on Gaussian elimination in which both upper and lower bounds in (3.1) are manipulated

simultaneously at each elimination and substitution stage. Such algorithms are easy to generate, but experience indicates that they often lead to bounding sets that are not maximally tight on S_k. This point has been noted in conjunction with the solution of interval linear equations in which both $\{X_{ij}\}$ and $\{y_i - e_i\}$ are permitted to assume any value in given intervals [9, p. 61]. Indeed, any algorithm for solving linear interval equations [9], [10] could be applied to (3.1), but here again these algorithms do not exploit the structure inherent in (3.1). In particular, it has been shown [10] that the exact solution set for the interval linear equation problem need not be convex, whereas the sets S_k considered here are.

The approach advocated here for computing O_k is to first determine the coordinates of the center $\mathbf{c}_0$ of the p-polytope S_k. Then, with $\mathbf{c}_0$ as the center of a local coordinate system, vectors $\mathbf{c}_1, \mathbf{c}_2, \cdots \mathbf{c}_p$ from $\mathbf{c}_0$ to the center of p nonparallel facets of S_k are determined. These vectors are then used, with the result of Theorem 3.1 and its corollary below, to construct the vertices at which the extremes of S_k occur. As will be seen from the results that follow, this approach reduces the problem of computing O_k to one of solving $p + 1$ sets of $p \times p$ simultaneous linear equations of the form $X\mathbf{c}_j = \mathbf{z}_j$.

Before giving a detailed pseudocode description of the algorithm just outlined, it is necessary to establish the following key results. First, note that S_k is a p-polytope with 2^p vertices

$$\mathbf{s}_j = [s_j^1, s_j^2, \cdots, s_j^p]^T$$

satisfying the linear equations

$$\sum_{i=1}^{p} X_{mi} s_j^i = t_j^m \tag{3.3}$$

for $m = 1, 2, \cdots, p$ and $j = 1, 2, \cdots, 2^p$. Here, $t_j^m = u_m$ or v_m for all j, and it will simplify the proof of Theorem 3.1 below to write this as

$$t_j^m = f_{jm} v_m + (1 - f_{jm}) u_m \tag{3.4}$$

where $f_{jm} = 0$ or 1. Similarly, the centers $\mathbf{c}_0, \mathbf{c}_1, \cdots, \mathbf{c}_p$ of S_k and its p principal facets are given by $\mathbf{c}_j = [c_j^1, c_j^2, \cdots, c_j^p]^T$, where

$$\sum_{i=1}^{p} X_{mi} c_j^i = z_j^m, \tag{3.5}$$

$$z_0^m = (u_m + v_m)/2, \tag{3.6a}$$

$$z_j^m = \begin{cases} (v_j - u_j)/2 & \text{if } m = j, \\ 0 & \text{otherwise} \end{cases} \tag{3.6b}$$

for $j = 1, 2, \cdots, p$. Note that the vector $\mathbf{z}^0 = [z_1^0, z_2^0, \cdots, z_p^0]^T$ defines the center of the constraint set, placing $\mathbf{c}_0$ in the center of S_k, equidistant from the parallel supporting hyperplanes defined by lower bound u_j and upper bound v_j, for all j. Similarly, the vector $\mathbf{z}^m = [z_1^m, z_2^m, \cdots, z_p^m]^T$ defines a vector $\mathbf{c}_m$ directed from $\mathbf{c}_0$ to the center of facet m, parallel to the other facets of S_k.

Given these results, the keys to obtaining an exact solution for the tightest bounding set O_k for (3.1) are the following theorem and its corollary.

THEOREM 3.1. *Suppose the matrix $X = [X_{ij}]$ has rank p. Then, any vertex s_j of S_k may be expressed as the vector sum*

$$\mathbf{s}_j = \mathbf{c}_0 + \sum_{r=1}^{p} [2f_{jr} - 1]\mathbf{c}_r. \tag{3.7}$$

Proof. Since X is full rank, the solutions of (3.3) and (3.5) are unique. Thus, it suffices to show that, given coefficients f_{jm} defining $\mathbf{s}_j$, (3.7) is consistent with (3.3)–(3.6).

Taking (3.7) component by component, we can expand the left-hand side of (3.3) as follows:

$$\sum_{r=1}^{p} X_{mr} s_j^r = z_0^m + \sum_{r=1}^{p} [2f_{jr} - 1] z_r^m$$

where (3.5) has been used to simplify the right-hand side. From (3.6), it follows that

$$\begin{aligned} z_0^m + \sum_{r=1}^{p} [2f_{jr} - 1] z_r^m &= (v_m + u_m)/2 + [2f_{jm} - 1](v_m - u_m)/2 \\ &= f_{jm} v_m + [1 - f_{jm}] u_m \\ &= t_j^m. \end{aligned}$$

Thus, (3.3) is satisfied, establishing the result. □

COROLLARY 1. *Given m, the maximum value of the mth component of any parameter vector in S_k is given by*

$$a_m^+ = c_0^m + \sigma_m \tag{3.8a}$$

and the minimum value is given by

$$a_m^- = c_0^m - \sigma_m \tag{3.8b}$$

where

$$\sigma_m = \sum_{r=1}^{p} |c_r^m|. \tag{3.9}$$

Proof. From Theorem 3.1, the mth component of any vertex vector $\mathbf{s}_j$ is given by

$$s_j^m = c_0^m + \sum_{r=1}^{p} q_{jr} c_r^m$$

where $q_{jr} = \pm 1$. Clearly, this sum is maximized when all terms are positive, corresponding to $q_{jr} = \operatorname{sgn}[c_r^m]$, from which (3.8a) follows immediately. Similarly, this sum is minimized when all terms are negative, from which (3.8b) follows by a similar argument. □

Given these results, an outer bounding algorithm may be implemented as follows.

set $O_0^* =$ a priori bounding set, if available;
otherwise, set $O_0^* = R^p$;

partition the available data into K subsets, either as in (3.2a), (3.2b), and (3.2c), or by using one of the schemes described in § 5;

do for $k = 1,2, \cdots ,K$

  do for $q = 0,1,2, \cdots ,p$

    —form the vector $\mathbf{z}_q$ from (3.6a), (3.6b)
    —solve $X\mathbf{c}_q = \mathbf{z}_q$ for $\mathbf{c}_q$

  end do

compute $O_k^{\#} = [a_1^-, a_1^+] \times \cdots \times [a_p^-, a_p^+]$ from (3.8a), (3.8b), and (3.9)

compute $O_k^* = O_k^{\#} \cap O_{k-1}^*$

end do

Algorithm returns O_K^* as the outer bounding set O^*.

4. Tightness of bounds O_k. A useful adjunct to the computational algorithm just described is an estimate of the tightness ω_k of O_k as a bound on S_k. It is easy to develop such a bound by considering the largest *inner bound* I_k contained in S_k of the same shape as O_k. That is, if $O_k = C(\mathbf{c}_0, \Sigma)$, take $I_k(h) = C(\mathbf{c}_0, h\Sigma)$ for the maximum h such that $O_k \supseteq S_k \supseteq I_k(h)$. It then follows that

$$\text{vol}\{O_k\} \geqq \text{vol}\{S_k\} \geqq \text{vol}\{I_k(h)\} = h^p \,\text{vol}\{O_k\}. \tag{4.1}$$

Thus, from the definition (2.2) of ω_k, the tightness of the outer bound O_k satisfies the inequality

$$1 \leqq \omega_k \leqq h^{-p}. \tag{4.2}$$

To compute h, it is most convenient to first change coordinates in a way that transforms the outer bounding set O_k into the unit cube, centered at the origin. Thus, we define centered and scaled parameters b_j by

$$b_j = (a_j - c_0^j)/\sigma_j \tag{4.3}$$

from which it follows by (3.8a, b) and (3.9) that

$$-1 \leqq b_j \leqq 1$$

for $j = 1, 2, \cdots, p$. The corresponding transformed values for X_{ij}, u_i, and v_i are, respectively,

$$X_{ij} \rightarrow F_{ij} = X_{ij}\sigma_j, \tag{4.4a}$$

$$u_i \rightarrow -g_i = (u_i - v_i)/2, \tag{4.4b}$$

$$v_i \rightarrow +g_i = (v_i - u_i)/2. \tag{4.4c}$$

Under this transformation, (3.1) defining S_k becomes

$$-g_i \leqq \sum_{j=1}^{p} F_{ij} b_j \leqq g_i \tag{4.5}$$

and the inner bound sought here is the largest cube

$$I_k(h) = [-h, h] \times \cdots \times [-h, h] \tag{4.6}$$

such that $S_k \supseteq I_k(h)$. This value of h is given by the following theorem.

THEOREM 4.1. *The largest h such that $S_k \supseteq I_k(h)$ is given by*

$$h = \min_i \left\{ g_i \Big/ \sum_{j=1}^{p} |F_{ij}| \right\}. \tag{4.7}$$

Proof. Here, $I_k(h)$ is the convex hull of the 2^p vertices

$$\mathbf{v}_r = [q_{1r}h, q_{2r}h, \cdots, q_{pr}h]^T$$

where $q_{jr} = \pm 1$ for $j = 1, 2, \cdots, p$ and $r = 1, 2, \cdots, 2^p$.

Thus, since S_k is convex, $S_k \supseteq I_k(h)$ is equivalent to $\mathbf{v}_r \in S_k$ for all r

$$\Rightarrow -g_i \leqq \sum_{j=1}^{p} F_{ij} q_{jr} h \leqq g_i$$

for all i, r. Since $q_{jr} = \pm 1$ for all j, r, it follows that

$$-\sum_{j=1}^{p} |F_{ij}| h \leqq \sum_{j=1}^{p} F_{ij} q_{jr} h \leqq \sum_{j=1}^{p} |F_{ij}| h$$

where both extremes are achieved by vertices of $I_k(h)$. Thus, $S_k \supseteq I_k(h)$ if and only if

$$\sum_{j=1}^{p} |F_{ij}| h \leqq g_i$$

for $i = 1, 2, \cdots, p$. To satisfy all of these constraints simultaneously, it follows that the maximum allowable value of h is given by the bound (4.7), as claimed. □

Geometrically, h represents a measure of the extent to which the facets of the polytope S_k lie parallel to the facets of the hypercube O_k. Consequently, if h is consistently very small for most of the K data partitions, this may be an indication that some of the parameters $\{a_j\}$ in the model under consideration are strongly coupled. If so, it may be desirable to identify and combine these coupled parameters to obtain a new, possibly simpler, model in which the free parameters are less strongly coupled and thus more clearly identifiable from the available data.

5. Data partitioning alternatives. The key step in developing the algorithm outlined in § 3 was the partitioning of the $N \times p$ matrix G into K $p \times p$ submatrices $\{X_k\}$. While a disjoint partitioning of the data was assumed for simplicity in § 3, the only essential requirement was that the resulting matrices X_k be nonsingular. A necessary condition on the selection of such nonsingular data partitionings is the following: Define a *nonredundant partitioning* $P_K(G)$ as any collection $\{X_k\}$ of K $p \times p$ matrices formed from the rows of G such that the same row does not appear more than once in any X_k. If all of the available data is to be used (i.e., if all rows of G are to be used at least once in the estimation process), it follows that nonredundant partitionings $P_K(G)$ exist only for $N/p \leqq K \leqq N!/(N-p)!p!$.

The case $K = N/p$ corresponds to the disjoint partitioning assumed in § 3, while the case $K = N!/(N-p)!p!$ is an impractically complex strategy that guarantees a complete vertex search of the polytope S^*. Consequently, this latter partitioning is guaranteed to yield the maximally tight bounding set O^* obtained by the linear programming algorithm of Milanese and Belforte [1], but with a lot more work in all but the most perverse examples. The key point is that the size K of the nonredundant partitioning used in computing the bounding set O_K^* provides a useful algorithm tuning parameter for trading off between computational efficiency and tightness of the resulting bounds. As the numerical example discussed in § 6.1 illustrates, a reasonable compromise appears to be the "sliding block" partitioning obtained for $K = N - p$ by combining each observation j (i.e., the jth row of G) with its $p - 1$ predecessors (i.e., rows $j - 1$ through $j - p + 1$). Besides representing a reasonable efficiency tradeoff of computational effort and estimation, this particular scheme is also suitable for real-time applications, yielding a new parameter bounding set O_k each time a new observation is obtained.

While nonredundancy is a generic necessary condition for the matrices $\{X_k\}$ to be nonsingular, sufficient conditions will be problem specific, as the following three examples

illustrate. First, consider the problem of fitting (x_i, y_i) coordinate pairs with a $(p-1)$th order polynomial basis set $\{p_j(x)\}$. In this case, model equation (1.1) becomes

$$y_i = \sum_{j=1}^{p} a_j p_{j-1}(x_i) + e_i, \tag{5.1}$$

so the G matrix elements are $G_{ij} = p_{j-1}(x_i)$. If the sequence $\{x_i\}$ is monotonically increasing, then a sufficient condition for any nonredundant partitioning to yield nonsingular matrices X_k is the Haar condition [11, p. 337] that every polynomial basis function $p_j(x)$ have at most $p-1$ zeros on the interval $[x_1, x_N]$.

Next, consider the problem of fitting a second-order moving average (MA) model to input-output data pairs (x_i, y_i). Here, model equation (1.1) becomes

$$y_i = a_1 x_i + a_2 x_{i-1} + e_i, \tag{5.2}$$

so the ith row of the G matrix is $[x_i, x_{i-1}]$. Nonsingularity of a matrix X_k formed by combining observations i and m requires the determinant $x_i x_{m-1} - x_{i-1} x_m$ to be nonzero. Practically, this requirement means that exponential input sequences $x_i = cr^i$ are not "persistently exciting" and do not exercise the system enough to identify both of the unknown parameters a_1 and a_2.

Similarly, consider the problem of fitting a second-order autoregressive (AR) model to a sequence $\{y_i\}$ of observed system responses. In this case, model equation (1.1) becomes

$$y_i = a_1 y_{i-1} + a_2 y_{i-2} + e_i, \tag{5.3}$$

yielding a G matrix with rows $[y_{i-1}, y_{i-2}]$. As in the MA model just considered, nonsingularity of the matrices X_k excludes exponential sequences $y_i = cr^i$. In this case, the assumed model is over-parameterized, since the exponential response is exactly describable by a first-order AR model.

Finally, another important consideration in selecting a data partitioning scheme is the numerical conditioning of the resulting matrices $\{X_k\}$. That is, even if they all yield nonsingular X_k matrices, different nonredundant data partitionings can exhibit very different behavior in practice due to differences in the singular value distributions of these matrices. As the numerical example in § 6.2 illustrates, both the estimation efficiency of the bounding set O^* and the computational effort required to determine it can vary drastically with different data partitioning schemes.

6. Numerical examples. To illustrate the performance of the computational algorithm proposed in § 3, this section presents two numerical examples. The first is an impractically simple geometric problem, but one that clearly illustrates the mechanics of all of the parallelepiped estimation algorithms considered here. The second example considers the very practical problem of evaluating exponential decay parameters from quantized experimental data.

6.1. Geometric example. Consider the following four inequalities:

$$0 \leqq a_1 + a_2 \leqq 2, \tag{6.1a}$$

$$0 \leqq -a_1 + a_2 \leqq 2, \tag{6.1b}$$

$$2 \leqq a_1 + 2a_2 \leqq 6, \tag{6.1c}$$

$$2 \leqq 2a_1 + a_2 \leqq 4. \tag{6.1d}$$

Here, the exact polytope S^* defined by these inequalities is a triangle with vertices $\{(\frac{2}{3}, \frac{2}{3}), (1, 1), (0, 2)\}$ lying between the hyperplanes defined by the upper bound of (6.1a) and the lower bounds of (6.1b) and (6.1d). The maximally tight bounding set that would be obtained from Milanese and Belforte's algorithm may be constructed by inspection from S^* as $B^* = [0, 1] \times [\frac{2}{3}, 2]$, shown in Fig. 1.

To apply the parallelepiped algorithm of Fogel and Huang, it is necessary to specify a priori bounds on a_1 and a_2. Their algorithm then solves (6.1a) together with the a priori bound on a_1 to obtain an a posteriori bound on a_2. Similarly, the a priori bound on a_2 is substituted into (6.1a) to obtain an a posteriori bound on a_1. This process is repeated for inequalities (6.1b), (6.1c), and (6.1d), using the a posteriori bound obtained at each stage as an a priori bound for the succeeding stage. If we assume $a_1 \in [-2, 2]$ and $a_2 \in [-5, 5]$ as a priori bounds, the final result of this process is the outer bound $B_{\text{FH1}} = [0, 2] \times [0, 3]$, a set 4.5 times larger in area than the minimal rectangle B^*. To see the dependence of the Fogel–Huang parallelepiped bounds on a priori estimates, note that if the starting estimate is $[-20, 20] \times [-20, 20]$, the resulting outer bounds are $B_{\text{FH2}} = [-5, \frac{15}{2}] \times [-11, 12]$, better than the a priori estimate, but over 200 times larger in area than B^*. This dependence on the quality of a priori parameter estimates is a key point since poor or nonexistent estimates must often be tolerated in practice.

For the bounding algorithm proposed in § 3, three possible sequential data partitions may be considered—inequalities (6.1a,b), (6.1b,c), and (6.1c,d). With these partitions numbered 1, 2, and 3, the corresponding bounding sets are

$$O_1 = [-1, 1] \times [0, 2], \quad O_2 = [-\tfrac{2}{3}, 2] \times [\tfrac{2}{3}, \tfrac{8}{3}], \quad O_3 = [-\tfrac{2}{3}, 2] \times [0, \tfrac{10}{3}].$$

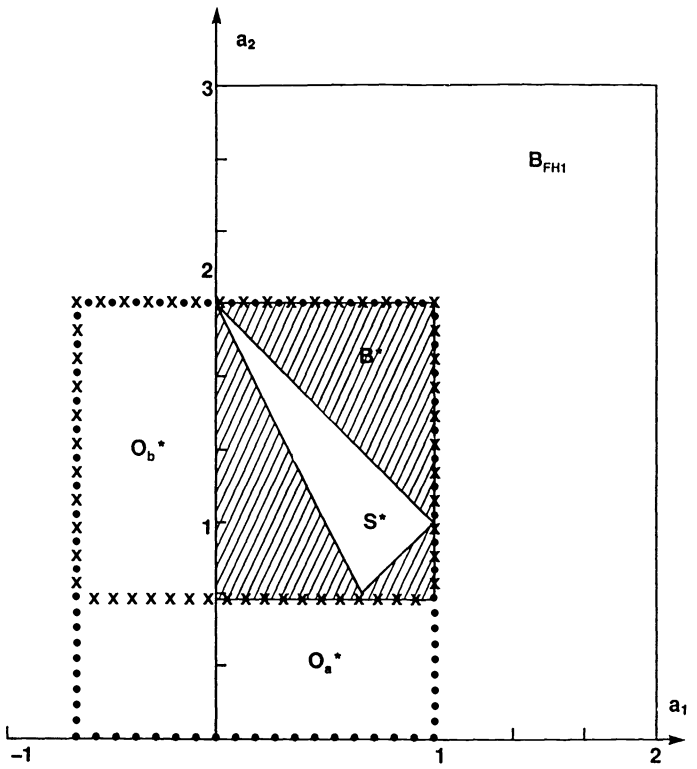


FIG. 1. *Estimate sets for geometric example.*

TABLE 1
Relative areas of parameter estimate sets.

Method	Set	Relative area
Exact solution	S^*	1.00
Milanese–Belforte	B^*	4.00
Fogel–Huang I	B_{FH1}	18.00
Fogel–Huang II	B_{FH2}	862.50
Section 3, $K = 2$	O_a^*	10.00
Section 3, $K = 3$	O_b^*	6.67

These bounds may be combined to obtain either

$$O_a^* = O_1 \cap O_3 = [-\tfrac{2}{3}, 1] \times [0, 2]$$

or

$$O_b^* = O_1 \cap O_2 \cap O_3 = [-\tfrac{2}{3}, 1] \times [\tfrac{2}{3}, 2].$$

For comparison, the relative areas of each of these bounding sets are listed in Table 1. It is clear that, for this simple example, the algorithm proposed in § 3 yields vastly improved parameter estimates relative to the Fogel–Huang parallelepiped algorithm, while retaining its advantages of sequential data processing. Also, note the significant improvement in performance for this example in going from $K = 2$ (2.5 times the Milanese–Belforte bounding set area) to $K = 3$ (1.67 times this area).

6.2. Quantized exponential decay example. As a more practical example, consider the identification of exponential decay parameters from empirical data. This is, suppose some physical measurand $y(t)$ is given by

$$y(t) = A \exp\{-t/T\} \tag{6.3}$$

for $t \geqq 0$. The time constant T is usually a quantity of significant interest, directly related to the physical relaxation phenomenon under study. Thus, a common experimental objective is to determine the unknown parameters A and T from discrete measurements of $y(t)$ taken at times $t_k = kT_s$ for $k = 0, 1, 2, \cdots, N-1$, following some experimentally applied stimulus at $t = 0$. The available data from which these parameters are to be estimated is the set of observations $\{y_k\}$, given by

$$y_k = A \exp\{-bk\} + e_k \tag{6.4}$$

where $b = T_s/T$ and e_k represents the observation error inherent in each measurement.

If the data is acquired by an m-bit A/D converter, the discrete observations $\{y_k\}$ will be related to the continuous-time signal $y(t)$ by

$$y_k = \begin{cases} 0, & y(t_k) < Q, \\ (j-1)Q, & (j-1)Q \leqq y(t_k) < jQ, \\ F, & y(t_k) \geqq F. \end{cases} \tag{6.5}$$

Here, $Q = 2^{-m}F$ is the quantization level (i.e., the "least significant bit") of the A/D converter, and $[0, F]$ represents its full-scale input range. Thus, if the error sequence $\{e_k\}$ in (6.4) represents quantization error $y_k - y(t_k)$, it will satisfy the set-theoretic constraint

$$-Q \leqq e_k \leqq 0 \tag{6.6}$$

for all k, provided $y(t)$ is restricted to the range $[Q, F)$.

In its present form, this problem is not linear in the unknown parameters, as required for the set-theoretic estimation strategy considered here. Equation (6.4) may be linearized, however, through the following manipulations. First, introduce a *relative* error r_k, defined as

$$r_k = e_k / y_k \tag{6.7}$$

and reduce the model (6.4) to

$$y_k(1 - r_k) = A \exp\{-bk\} \tag{6.8}$$

where r_k satisfies the set-theoretic error constraint

$$-Q/y_k \leqq r_k \leqq 0. \tag{6.9}$$

Equation (6.8) may then be linearized by taking logarithms of both sides, yielding

$$z_k = a - bk + h_k \tag{6.10}$$

where

$$z_k = \ln y_k, \tag{6.11a}$$

$$a = \ln A, \tag{6.11b}$$

$$h_k = -\ln(1 - r_k) \tag{6.11c}$$

and the error h_k satisfies the bounds

$$-L_k \leqq h_k \leqq 0 \tag{6.12a}$$

for

$$L_k = \ln(1 + Q/y_k). \tag{6.12b}$$

Applying the data partitioning introduced in § 3 to the logarithmically transformed data $\{z_k\}$ and $\{L_k\}$ yields $N/2$ pairs of inequalities of the form

$$z_{2k-2} \leqq a - (2k-2)b \leqq z_{2k-2} + L_{2k-2}, \tag{6.13a}$$

$$z_{2k-1} \leqq a - (2k-1)b \leqq z_{2k-1} + L_{2k-1}, \tag{6.13b}$$

for $k = 1, 2, \cdots, N/2$. Due to the simplicity of these inequalities, analytic expressions may be developed for the bounds of § 3. Denoting the lower bounds in (6.13a) and (6.13b) as u_1 and u_2, respectively, and the corresponding upper bounds as v_1 and v_2, the parameter bounds become

$$a_1^- = (2k-1)u_1 - (2k-2)v_2, \tag{6.14a}$$

$$a_1^+ = (2k-1)v_1 - (2k-2)u_2, \tag{6.14b}$$

$$a_2^- = u_1 - v_2, \tag{6.14c}$$

$$a_2^+ = v_1 - u_2. \tag{6.14d}$$

Numerical values for these bounds are given in Table 2 for $A = 1$ (implying $a_1 = 0$) and $b = 0.07$. All of these values were computed from a 100-point simulation of (6.4) with $e_k = 0$ in single precision FORTRAN-77 on a VAX-11/785 computer. These values were then quantized according to (6.5) with $F = 1$ and $m = 8$, 12, or 16. Whenever this process resulted in a quantized value of zero, the data set was truncated to include only nonzero data points. This problem only occurred with the eight-bit quan-

TABLE 2
Exponential decay parameter bounds—uniformly sampled case.

m	a_1^-	a_1^+	a_2^-	a_2^+
8	0.0000	0.00390	0.06514	0.07261
12	0.0000	0.00244	0.06991	0.07022
16	0.0000	0.00015	0.06999	0.07004
Exact:	$a_1 = 0.00000$		$a_2 = 0.0700$	

tization, reducing the quantized data set from 100 points to 90 points. The results given in Table 2 represent the final bounding intervals returned by the algorithm described in § 3.

An examination of these results shows that, as expected, the bounding intervals shrink toward the exact parameter values as m increases, with the difference between the eight-bit and the 16-bit results being fairly dramatic. Contrary to expectations, however, these bounding intervals were determined by the first few data points alone, so the estimate sets *do not* converge to the corresponding point estimates as $N \to \infty$. Examination of the parameter bounding intervals computed at each step k individually shows that these intervals become wider with increasing k and thus do not influence the final intersection that defines the overall outer bound O^*.

This phenomenon is a reflection of two facets of the exponential decay problem considered here. First, note that while the bounds on the absolute quantization error e_k are constant, the relative error bounds grow because the measured signal y_k is monotonically decreasing. Thus, it is to be expected that the parameter uncertainty intervals will grow with increasing k. More serious, however, is the fact that the data matrix used at step k of the estimation process is

$$X_k = \begin{bmatrix} 1 & 2k-2 \\ 1 & 2k-1 \end{bmatrix} \tag{6.15}$$

which becomes increasingly ill-conditioned with increasing k.

Another indication of the ill-conditioning of X is the behavior of the scaling parameter h introduced in § 4. If we define the error growth ratio R as

$$R = (v_2 - u_2)/(v_1 - u_1) = L_{2k-1}/L_{2k-2}, \tag{6.16}$$

this scaling parameter is the smaller of

$$h_1 = 1/[(4k-3)+(4k-4)R] \tag{6.17a}$$

and

$$h_2 = 1/[(4k-3)+(4k-2)/R]. \tag{6.17b}$$

In the eight-bit case, h diminishes from approximately 0.53549 for $k = 1$ to approximately 0.00282 for $k = 45$, correctly indicating that the bounds become quite loose for large k.

To improve the conditioning of the X matrix, an alternative experimental design was developed. Specifically, rather than sampling the data uniformly at times $t_n = nT_s$, measurements were assumed to be taken at instants $t_n = 2^n T_s$. This measurement strategy changed the data matrix X_k from that given by (6.15) to

$$X_k = \begin{bmatrix} 1 & 2^{2k-2} \\ 1 & 2^{2k-1} \end{bmatrix}. \tag{6.18}$$

TABLE 3
Parameter bounds for exponential decay—exponentially sampled case.

m	a_1^-	a_1^+	a_2^-	a_2^+
8	−0.00388	0.00010	0.06866	0.07106
12	−0.00035	0.00027	0.06995	0.07003
16	−0.00003	0.00000	0.07000	0.07000
Exact:	$a_1 = 0.00000$		$a_2 = 0.07000$	

The conditioning of this data matrix also becomes worse with increasing k, but much more slowly than the data matrix (6.15). Bounding intervals for this sampling strategy are given in Table 3, which shows them to be generally better than those obtained in the uniformly sampled case. This is an extremely significant conclusion, since these bounds were computed from much less data, e.g., three data partitions (six inequalities) versus 45 or 50 data partitions (90 or 100 inequalities). Also, note that in contrast to the uniformly sampled case, the scaling parameters h_1 and h_2 in (6.17a,b) depend only weakly on k in the exponentially sampled case (through R's dependence on k) and are given by

$$h_1 = 1/[3+2R] \tag{6.19a}$$

and

$$h_2 = 1/[3+4/R]. \tag{6.19b}$$

These parameters are consistently larger than in the uniformly sampled problem, remaining on the order of 0.1 for all of the cases considered here.

7. Summary. This paper has described a new parallelepiped-based outer bounding algorithm for the set-theoretic estimation problem. While the bounds obtained are not as tight as those of Milanese and Belforte, they are considerably easier to evaluate. In addition, because it processes the data sequentially, this algorithm can be used in applications requiring real-time parameter updates or adaptive parameter estimation. Because of its ease of implementation, especially for small p, and the fact that it yields independent parameter bounds, this outer bounding algorithm should serve as an extremely useful tool in preliminary data analysis. For example, it could be used in place of Fogel and Huang's ellipsoidal algorithm [8] in Belforte, Bona, and Cerone's scheme [5] of preprocessing the data to reduce the number of active constraint candidates in Milanese and Belforte's linear programming method [1]. Similarly, the algorithm proposed here could be used as a general prescreening tool to identify regions in the model parameter space that are worthy of greater scrutiny by other methods.

Finally, because they are not nearly as well known as statistical confidence interval estimators, it is probably appropriate to conclude with a few remarks concerning the philosophical and practical implications of set-theoretic estimators. First, note that in spite of the apparent similarities between set-theoretic estimation and statistical parameter estimation with uniform error distributions, the two are *not* equivalent. In particular, the uniform error distribution assumption is stronger than the set-theoretic error constraint. If $e_i^- = e^-$ and $e_i^+ = e^+$ for all i, then the exact set-theoretic estimate set S^* does represent the 100-percent confidence set in R^p for the parameters, estimated under an assumption of independently and identically distributed measurement errors, uniformly distributed on $[e^-, e^+]$. However, S^* may equally well be viewed as the 100-percent confidence set for *any* independently and identically distributed measurement error distribution of compact support on $[e^-, e^+]$-skewed, multimodal, discrete, etc. In fact,

deterministic error models describing bounded amplitude sinusoidal oscillations, polynomial drifts over finite observation times, or chaotic processes are also consistent with the set-theoretic description. Thus, without additional information regarding this error distribution, the set S^* should not be viewed as anything other than a set of worst-case parameter bounds. In particular, without such additional information, there is no reason to select any single point within S^* as a preferred point estimator of the unknown parameter vector in R^p. Conversely, if statistical error models are appropriate, available, and computationally tractable, they are to be preferred, since they provide much more information about the unknown parameters (i.e., both defensible point estimates and uncertainty intervals). In many applications, however, these conditions are not satisfied, implying a need for practical approaches like the one described here for obtaining parameter uncertainty estimates from measured data.

Acknowledgments. The author acknowledges enlightening discussions and correspondence with E. Fogel, Y. F. Huang, G. Belforte, B. Bono, and G. C. Verghese at various stages in the development of the ideas described here. Thanks also go to the referees of the earlier drafts of this manuscript for their valuable insights and useful recommendations.

REFERENCES

[1] M. MILANESE AND G. BELFORTE, *Estimation theory and uncertainty intervals evaluation in presence of unknown but bounded errors: linear families of models and estimators*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 408–414.

[2] F. C. SCHWEPPE, *Uncertain Dynamic Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[3] B. GRUNBAUM, *Convex Polytopes*, John Wiley, New York, 1967.

[4] W. KAHAN, *Circumscribing an ellipsoid about the intersection of two ellipsoids*, Canad. Math. Bull., 11 (1968), pp. 437–441.

[5] G. BELFORTE, B. BONA, AND V. CERONE, *An improved parameter identification algorithm for signals with unknown-but-bounded errors*, in Proc. 7th IFAC Symposium on Identification and System Parameter Estimation, York, United Kingdom, 1985.

[6] E. FOGEL AND Y. F. HUANG, *Adaptive algorithms for non-statistical parameter estimation in linear models*, in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Denver, Colorado, 1980, pp. 1022–1025.

[7] S. MACLANE AND G. BIRKHOFF, *Algebra*, Macmillan, New York, 1979.

[8] E. FOGEL AND Y. F. HUANG, *On the value of information in system identification-bounded noise case*, Automatica, 18 (1982), pp. 229–238.

[9] R. E. MOORE, *Methods and Applications of Interval Analysis*, Society for Industrial and Applied Mathematics, Philadelphia, 1979.

[10] D. M. GAY, *Solving interval linear equations*, SIAM J. Numer. Anal., 19 (1982), pp. 858–870.

[11] E. KREYSZIG, *Introductory Functional Analysis with Applications*, John Wiley, New York, 1978.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 4, October 1988

006

HOMOTOPY METHOD FOR GENERAL λ-MATRIX PROBLEMS*

MOODY T. CHU†, T. Y. LI‡, AND TIM SAUER§

Abstract. This paper describes a homotopy method used to solve the kth-degree λ-matrix problem $(A_k\lambda^k + A_{k-1}\lambda^{k-1} + \cdots + A_1\lambda + A_0)x = 0$. A special homotopy equation is constructed for the case where all coefficients are general $n \times n$ complex matrices. Smooth curves connecting trivial solutions to desired eigenpairs are shown to exist. The homotopy equations maintain the nonzero structure of the underlying matrices (if there is any) and the curves correspond only to different initial values of the same ordinary differential equation. Therefore, the method might be used to find all isolated eigenpairs for large-scale λ-matrix problems on single-instruction multiple data (SIMD) machines.

Key words. λ-matrix, homotopy continuation method, zeros of polynomial systems

AMS(MOS) subject classifications. 65H10, 58C99, 55M25

1. Introduction.

Given a kth-degree matrix polynomial

$$P(\lambda) = A_k\lambda^k + A_{k-1}\lambda^{k-1} + \cdots + A_1\lambda + A_0 \tag{1.1}$$

with $A_k, A_{k-1}, \cdots, A_0 \in \mathbb{C}^{n\times n}$, the λ-matrix problem consists of determining scalars λ, called eigenvalues, and corresponding $n \times 1$ nonzero vectors x, called eigenvectors, such that

$$P(\lambda)x = 0 \tag{1.2}$$

is satisfied. Problems of this kind occur in many different application areas. Note that the important regular eigenvalue problem

$$\lambda x = Ax \tag{1.3}$$

and the generalized eigenvalue problem

$$\lambda Bx = Ax \tag{1.4}$$

are just two special linear cases of the general problem (1.2). Various examples of (1.1) in physical applications can be found in [3]–[5] and the references cited therein.

A variety of numerical methods are available for solving the λ-matrix problem. In fact, several review papers have already appeared. Without attempting a complete list, we mention here only those by Gohberg, Lancaster, and Rodman [3], Lancaster [4], [5], Ruhe [12], Scott [13], and Peters and Wilkinson [10]. Roughly, most of the approaches can be classified into three categories:

(1) Solving the linearized problem;
(2) Iterating directly;
(3) Reducing to the canonical form.

Each approach has its strengths and weaknesses. For example, the first approach can make use of the readily available software packages, but it increases the size considerably. The second approach includes subspace and Newton-type iterations. Both iterative pro-

* Received by the editors October 26, 1987; accepted for publication (in revised form) March 30, 1988.
† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.
‡ Department of Mathematics, Michigan State University, East Lansing, Michigan 48824. The research of this author was supported in part by the Defense Advanced Research Projects Agency, and in part by National Science Foundation grant DMS-8416503.
§ Department of Mathematics, George Mason University, Fairfax, Virginia 22030. The research of this author was supported in part by the Defense Advanced Research Projects Agency.

cesses are plausible in theory. However, concerns over the rate of convergence for the former process and the starting procedure for the latter arise in practice. The third approach involves the problem of finding zeros of one-dimensional polynomials. When the degree increases, this becomes an ill-conditioned problem. Interested readers may find more detailed discussions and references concerning each approach among the review papers mentioned above.

Recently the homotopy method has been applied successfully to find all isolated solutions of the linear algebraic eigenvalue problems. In [1], Chu proposes a homotopy equation for (1.3) when A is real, symmetric, and tridiagonal with nonzero off-diagonal elements. Li and Sauer [7] and Li, Sauer, and Yorke [8] study homotopy methods for (1.3) and (1.4) by using fairly sophisticated concepts from algebraic geometry when both A and B are general matrices. In [2] Chu shows that the equation formed in [1] for tridiagonal symmetric matrices works equally well for general matrices by using elementary algebraic theory. The same idea is also applicable to problem (1.4).

Solving the λ-matrix problem by the homotopy method may be costly because of the task of following the homotopy curves. We feel that with improvements in the curve-tracing techniques (say, a hybrid method) this overhead would be substantially reduced. Recently, Rhee [11] has reported some rather promising results on this subject. On the other hand, the homotopy method may have the following advantages:

(1) All isolated eigenpairs are guaranteed to be reached. The method can even approximate nonisolated eigenpairs.

(2) The homotopy curves correspond only to different initial values of the same ordinary differential equation. Hence, all curves can be followed simultaneously if there are enough processors.

(3) The homotopy equation respects the matrix structure (if there is any) of the original problem.

In this paper we present a general treatment of the homotopy method for solving the general kth-degree λ-matrix problem (1.2). Previous results in regard to the linear algebraic eigenvalue problems should then follow as special cases. Readers should be cautioned, however, that the line of thinking in this paper is fundamentally different from that of previous papers.

This paper is organized as follows. In § 2 we begin with a collection of preliminary observations. All these facts are either easy to prove or well known in the literature. We then use these fundamentally important facts to establish the theory of the homotopy method in § 3. Comments on computational aspects of our method are given in § 4, along with some numerical examples.

2. Preliminaries. In this section we observe some basic facts that will be used in the development of our homotopy method.

Consider an arbitrary λ-matrix

$$P(\lambda; B_k, \cdots, B_0) = B_k\lambda^k + \cdots + B_1\lambda + B_0, \tag{2.1}$$

where $B_k, \cdots, B_0 \in \mathbb{C}^{n\times n}$. When it becomes unambiguous, we shall abbreviate $P(\lambda; B_k, \cdots, B_0)$ as $P(\lambda)$.

We first observe the obvious fact that the determinant of $P(\lambda)$ is a polynomial. Indeed, $\det(P(\lambda)) = (\det(B_k))\lambda^{nk}$ + lower-degree terms. It follows, if we count the multiplicities, that the λ-matrix problem corresponding to (2.1) has exactly nk eigenvalues if the leading coefficient B_k is nonsingular. Such a λ-matrix is said to be regular.

Recall that the resultant $R = R(a_n, \cdots, a_0, b_m, \cdots, b_0)$ of two polynomials

$$f(x) = a_n x^n + \cdots + a_1 x + a_0,$$

$$g(x) = b_m x^m + \cdots + b_1 x + b_0,$$

with $a_n, \cdots, a_0, b_m, \cdots, b_0 \in \mathbb{C}$, $a_n \neq 0$, and $b_m \neq 0$ is defined to be the determinant of the $(n+m) \times (n+m)$ matrix

$$\begin{bmatrix} a_0, a_1, \cdots, a_n & & & \\ & a_0, a_1, \cdots, a_n & & \\ & & \cdots\cdots\cdots & \\ & & & a_0, a_1, \cdots, a_n \\ b_0, b_1, \cdots, b_m & & & \\ & b_0, b_1, \cdots, b_m & & \\ & & \cdots\cdots\cdots & \\ & & & b_0, b_1, \cdots, b_m \end{bmatrix},$$

which is made of m rows of a's, n rows of b's, and zeros elsewhere. It is well known [14] that f and g have a common nonconstant factor if and only if $R = 0$. Thus a polynomial f has a multiple root if and only if its discriminant, the resultant of f and its derivative f', is zero.

Given $d_i \in \mathbb{C}$, $i = 1, \cdots, n$, let $D = \operatorname{diag}(d_1, \cdots, d_n)$ and

$$p(\lambda) = \det(P(\lambda) - D). \tag{2.2}$$

We claim the following.

LEMMA 2.1. *There exist real numbers $(d_1, \cdots, d_n)$ such that $p(\lambda)$ has no multiple roots.*

Proof. We prove the lemma by induction on n, the size of $P(\lambda)$. For convenience, we rename the polynomial $p(\lambda)$ as $p_n(\lambda)$.

When $n = 1$, $p_1(\lambda)$ has multiple roots if and only if the discriminant $R_1(d_1)$ of $p_1(\lambda)$ vanishes. Suppose the leading coefficient of $p_1(\lambda)$ is a_k. It is easy to see that $R(d_1)$ is an $(n-1)$th polynomial in d_1 with leading coefficient $(ka_k)^k$. Therefore, $R(d_1)$ can vanish only at finitely many points. There exists a real number d_1 such that $p_1(\lambda)$ has no multiple roots.

Let $d_1, \cdots, d_{n-1}$ be chosen by the induction hypothesis so that $p_{n-1}(\lambda)$, the determinant of the principal $(n-1) \times (n-1)$ minor of $P(\lambda) - D$, has no multiple roots. With these fixed values of $d_1, \cdots, d_{n-1}$, we have

$$p_n(\lambda) = q_n(\lambda) - d_n p_{n-1}(\lambda)$$

where $q_n(\lambda)$ and $p_{n-1}(\lambda)$ do not depend upon the value of d_n. We claim the set of real-valued d_n such that $p_n(\lambda)$ has no multiple roots is dense in $\mathbb{R}$.

Suppose not. Then there would exist an open interval I such that $\lambda(d_n)$ is a multiple root of $p_n(\lambda) = p_n(\lambda; d_n)$. By refining the interval I if necessary, we may assume without loss of generality that $\lambda(d_n)$ is differentiable with respect to d_n. For each $d_n \in I$, we have

$$0 = p_n(\lambda(d_n)) = q_n(\lambda(d_n)) - d_n p_{n-1}(\lambda(d_n)).$$

Upon differentiating with respect to the parameter d_n, we get

$$\begin{aligned} 0 &= q_n'(\lambda(d_n))\lambda'(d_n) - d_n p_{n-1}'(\lambda(d_n))\lambda_n'(d_n) - p_{n-1}(\lambda(d_n)) \\ &= \left[\frac{d}{d\lambda} p_n(\lambda)\right]_{\lambda = \lambda(d_n)} \lambda'(d_n) - p_{n-1}(\lambda(d_n)) \\ &= -p_{n-1}(\lambda(d_n)), \qquad d_n \in I. \end{aligned}$$

The last equality follows from the fact that $\lambda(d_n)$ is a multiple root for $d_n \in I$. Note that $p_{n-1}(\lambda(d_n)) \equiv 0$ for $d_n \in I$ implies $\lambda(d_n) \equiv \lambda_0$ (a constant) for $d_n \in I$, since $p_{n-1}(\lambda)$ is a

polynomial. It follows that $p_n(\lambda)$ has a multiple root at $\lambda = \lambda_0$ for all $d_n \in I$. Choose $d_n^{(1)} \neq d_n^{(2)}$ in I. Then

$$0 = p_n(\lambda_0) = q_n(\lambda_0) - d_n^{(1)} p_{n-1}(\lambda_0),$$

$$0 = p_n(\lambda_0) = q_n(\lambda_0) - d_n^{(2)} p_{n-1}(\lambda_0),$$

$$0 = p_n'(\lambda_0) = q_n'(\lambda_0) - d_n^{(1)} p_{n-1}'(\lambda_0),$$

$$0 = p_n'(\lambda_0) = q_n'(\lambda_0) - d_n^{(2)} p_{n-1}'(\lambda_0).$$

It follows that $p_{n-1}(\lambda_0) = p_{n-1}'(\lambda_0) = 0$. This contradicts the induction hypothesis that $p_{n-1}(\lambda)$ has no multiple roots. □

The following lemma is an extension of the preceding result.

LEMMA 2.2. *The polynomial* $p(\lambda)$ *in* (2.2) *has no multiple roots for* $(d_1, \cdots, d_n)$ *almost everywhere in* $\mathbb{C}^n$ *except on a subset of complex codimension* 1.

Proof. The polynomial $p(\lambda)$ has no multiple roots if and only if its discriminant $R(d_1, \cdots, d_n)$ is nonzero. By Lemma 2.1, we know that $R(d_1, \cdots, d_n)$ is not identically zero. Furthermore, since $R(d_1, \cdots, d_n)$ is itself a polynomial in variables $d_1, \cdots, d_n$, it can vanish only on a hypersurface of complex codimension 1 [see 9]. □

It is well known in basic matrix theory that if all eigenvalues of a matrix are distinct, then it has no generalized eigenvectors. In [3] and [6], it is shown that this concept can be extended naturally to matrix polynomials. In particular, the following lemma is equivalent to the statement that there are no generalized eigenvectors [6, eq. 14.3.3] for the λ-matrix $P(\lambda)$. Readers are referred to [3, Chap. 1] and [6, Chap. 14] for more detailed discussions. We simply state the result without proof.

LEMMA 2.3. *Suppose the* λ*-matrix* $P(\lambda)$ *has* nk *distinct eigenvalues* $\lambda_1, \cdots, \lambda_{nk}$. *Let* x_j *be a unit eigenvector of* $P(\lambda)$ *associated with* λ_j, *i.e.*, $P(\lambda_j)x_j = 0$. *Then* $P'(\lambda_j)x_j \notin \text{Range}\,(P(\lambda_j))$, *where* $P'(\lambda) = (d/d\lambda)P(\lambda) = kB_k\lambda^{k-1} + \cdots + B_1$.

Henceforth we shall assume that the λ-matrix $P(\lambda)$ has nk distinct eigenvalues. For each eigenpair (x, λ) of $P(\lambda)$, we define $Q = Q(x, \lambda)$ to be the $n \times (n+1)$ complex matrix

$$Q(x, \lambda) = [P(\lambda), P'(\lambda)x]. \tag{2.3}$$

It follows from Lemma 2.3 that Q is of complex rank n.

Recall that a linear transformation from $\mathbb{C}^{n+1}$ to $\mathbb{C}^n$ can be regarded as a linear transformation from $\mathbb{R}^{2n+2}$ to $\mathbb{R}^{2n}$ if each component, say $z = a + ib$, of the complex matrix is replaced by the 2×2 real matrix $\left[\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right]$. Let $\hat{Q} \in \mathbb{R}^{2n \times (2n+2)}$ denote the real matrix associated with the complex matrix $Q \in \mathbb{C}^{n \times (n+1)}$ defined in (2.3). Suppose each component x_k of the complex vector x is written as $x_k = a_k + ib_k$, $k = 1, \cdots, n$. We define a matrix $M = M(x, \lambda) \in \mathbb{R}^{(2n+1) \times (2n+2)}$ as follows:

$$M(x, \lambda) = \begin{bmatrix} \hat{Q} \\ a_1, b_1, a_2, \cdots, a_n, b_n, 0, 0 \end{bmatrix}. \tag{2.4}$$

Note that the last row of M is orthogonal to all rows of $\hat{Q}$ because $P(\lambda)x = 0$. It follows that the matrix M is of real rank $2n + 1$.

3. Homotopy method. Equipped with the knowledge of the preceding section, we now consider our original λ-matrix problem (1.2).

For simplicity, we shall denote

$$P(\lambda)=A_k\lambda^k+A_{k-1}\lambda^{k-1}+\cdots+A_1\lambda+A_0, \tag{3.1}$$

$$Q(\lambda)=cI\lambda^k-D, \tag{3.2}$$

$$R(\lambda,t,c,D)=(1-t)Q(\lambda)+tP(\lambda) \tag{3.4}$$

where $c\in\mathbb{C}$ and $D=\operatorname{diag}(d_1,\cdots,d_n)\in\mathbb{C}^{n\times n}$ are to be specified later.

Observe that

$$\begin{aligned}R(\lambda,t,c,D)&=(1-t)Q(\lambda)+tP(\lambda)\\&=[(1-t)cI+tA_k]\lambda^k+tA_{k-1}\lambda^{k-1}+\cdots+tA_1\lambda+[(1-t)D+tA_0]\end{aligned}$$

is still a λ-matrix. It is easy to show [7] that there exists an open dense set U_1 with full measure in $\mathbb{C}$ such that if $c\in U_1$, then $[(1-t)cI+tA_k]$ is nonsingular for all $t\in[0,1)$. Henceforth we shall assume that the scalar c in (3.2) is always chosen from U_1, and abbreviate $R(\lambda,t,c,D)$ as $R(\lambda,t,D)$. We shall also denote

$$\begin{aligned}R_\lambda(\lambda,t)&=\frac{d}{d\lambda}R(\lambda,t,D)\\&=k[(1-t)cI+tA_k]\lambda^{k-1}+(k-1)tA_{k-1}\lambda^{k-2}+\cdots+tA_1.\end{aligned} \tag{3.4}$$

For the λ-matrix problem (1.2), the homotopy function $H\colon \mathbb{C}^n\times\mathbb{C}\times[0,1)\to\mathbb{C}^n\times\mathbb{R}$ is constructed as follows:

$$H(x,\lambda,t)=\begin{bmatrix}R(\lambda,t,D)x\\(x^*x-1)/2\end{bmatrix} \tag{3.5}$$

where x^* represents the transpose of the complex conjugate of x. We are interested in the set $H^{-1}(0)$. As our main result is we show that $H^{-1}(0)$ is a two-dimensional smooth submanifold in $\mathbb{R}^{2n}\times\mathbb{R}^2\times\mathbb{R}$.

Note first that $H(x,\lambda,1)=0$ corresponds to problem (1.2) with normalized eigenvectors. For $i=1,\cdots,n$, let e_i represent the standard ith unit vector in $\mathbb{R}^n$ and λ_{ij} be the jth complex root of $(d_i/c)^{1/k}$, where $j=1,\cdots,k$. It is obvious that $(e_i,\lambda_{ij},0)\in H^{-1}(0)$. We shall use these nk points $(e_i,\lambda_{ij},0)\in\mathbb{C}^n\times\mathbb{C}\times[0,1)$ as our initial points when constructing homotopy curves connected to the desired solution of (1.2).

The following theorem is the main result.

THEOREM 3.1. *There exists an open dense subset U of full measure in $\mathbb{C}^n$ such that, for $D=\operatorname{diag}(d_1,\cdots,d_n)$ with $(d_1,\cdots,d_n)\in U$ and each initial point $y_{ij}=(e_i,\lambda_{ij},0)$, the connected component $C(y_{ij})$ of y_{ij} in $H^{-1}(0)$, when identified as a subset in $\mathbb{R}^{2n}\times\mathbb{R}^2\times\mathbb{R}$, has the following properties.*

1. *$C(y_{ij})$ is a (real) analytic submanifold in $\mathbb{R}^{2n}\times\mathbb{R}^2\times\mathbb{R}$ with real dimension 2.*

2. *The cross-section of $C(y_{ij})$ with each hyperplane $t\equiv$ constant $\in[0,1)$ is a unit circle centered at $(0,\lambda)\in\mathbb{R}^{2n}\times\mathbb{R}^2$ for some λ.*

3. *The manifolds $C(y_{ij})$ corresponding to different initial points do not intersect for $t\in[0,1)$.*

4. *Each manifold $C(y_{ij})$ is bounded for $t\in[0,1)$.*

Proof. For each fixed $t\in[0,1)$, consider the λ-matrix

$$\bar{R}(\lambda,t,D)=\frac{1}{1-t}R(\lambda,t,D)=(cI\lambda^k-D)+\frac{t}{1-t}P(\lambda).$$

By Lemma 2.2, there exists a hypersurface $\bar{U}(t) \in \mathbb{C}^n$ of complex codimension 1 (real codimension 2) such that if $(d_1, \cdots, d_n) \notin \bar{U}(t)$, then $\det(\bar{R}(\lambda, t, D))$ does not have multiple roots. As t varies in $[0, 1)$, the set $V = \cup_{t\in[0,1)} \bar{U}(t) \subset \mathbb{C}^n$ is of real codimension at most one. Thus the complement U of V in $\mathbb{C}^n$ is open and dense and has full measure.

For $D = \operatorname{diag}(d_1, \cdots, d_n)$ with $(d_1, \cdots, d_n) \in U$, the λ-matrix $R(\lambda, t, D)$ has no multiple eigenvalues. For every $(x, \lambda, t) \in H^{-1}(0)$, it is necessary that $R(\lambda, t, D)x = 0$, i.e., x is an eigenvector of $R(\lambda, t, D)$ associated with the eigenvalue λ. Analogous to (2.3) we now consider the matrix $Q = Q(x, \lambda, t) \in \mathbb{C}^{n\times(n+1)}$, where

$$Q(x,\lambda,t) = [R(\lambda,t,D), R_\lambda(\lambda,t)x] \tag{3.6}$$

and its associated real matrix $M = M(x, \lambda, t) \in \mathbb{R}^{(2n+1)\times(2n+2)}$ is as defined in (2.4). Note that the homotopy function H may be regarded as a mapping from $\mathbb{R}^{2n} \times \mathbb{R}^2 \times \mathbb{R}$ into $\mathbb{R}^{2n} \times \mathbb{R}$ and that $M = (\partial H/\partial(x, \lambda))$, where the derivatives are taken in the real variable sense. By the way the constant $c \in \mathbb{C}$ and the vector $(d_1, \cdots, d_n) \in \mathbb{C}^n$ are selected, we know that the λ-matrix $R(\lambda, t, D)$ has nk distinct eigenvalues for every $t \in [0, 1)$. From the discussion in the preceding section, it follows that M is of full rank.

We may now apply the implicit function theorem to conclude that $H^{-1}(0)$ is a smooth submanifold in $\mathbb{R}^{2n} \times \mathbb{R}^2 \times \mathbb{R}$ with real dimension two. Assertion (1) is proved.

Indeed, given an arbitrary point $(x, \lambda, t) \in H^{-1}(0)$, note that the partial derivatives in forming the matrix M is not taken with respect to t. So a local neighborhood of (x, λ, t) on $H^{-1}(0)$ is diffeomorphic to a two-dimensional neighborhood of t and another suitable real variable from (x, λ). This shows that $H^{-1}(0)$ intersects each hyperplane $t \equiv$ constant $\in [0, 1)$ transversally. If the connected components $C(y_{i_1j_1})$ and $C(y_{i_2j_2})$ of two distinct initial points $y_{i_1j_1}$ and $y_{i_2j_2}$ ever intersect, then $C(y_{i_1j_1}) = C(y_{i_2j_2})$. This is possible only if at the intersection point the two-dimensional surface $C(y_{i_1j_1})$ "bends" back toward the initial point $y_{i_2j_2}$. This contradicts the transversal property we have observed. Assertion 3 is proved.

Since $H(x, \lambda, t) = 0$ implies $H(\gamma x, \lambda, t) = 0$ whenever $\gamma \in \mathbb{C}$ and $|\gamma| = 1$, we see that $C(y_{ij})$ indeed is a two-dimensional cylindrical tube whose cross-section with each hyperplane $t \equiv$ constant $\in [0, 1)$ is a unit circle centered at $(0, \lambda) \in \mathbb{R}^{2n} \times \mathbb{R}^2$ for some λ. Assertion 2 is proved.

To prove assertion 4 it remains to show only that on every manifold $C(y_{ij})$ the eigenvalue λ stays bounded for $t \in [0, 1)$. From assertions 2 and 3, it suffices to consider any one-dimensional submanifold on $C(y_{ij})$ that is parameterized by the variable t. Define

$$w = \min_{t\in[0,t_0]} \|[(1-t)cI + tA_k]x(t)\|.$$

Since $(1 - t)cI + tA_k$ is continuous and nonsingular for all $t \in [0, t_0]$, we have $w > 0$. Let $r(t) = |\lambda(t)|$ and $s = \max_{t\in[0,t_0]} \|tA_0 - (1 - t)D\|$. Then, $R(\lambda(t), t, D)x(t) = 0$ implies that

$$\begin{aligned} 0 < wr^k(t) &\leqq \|[(1-t)cI + tA_k]\lambda^k(t)\| \\ &\leqq t\|A_{k-1}\lambda^{k-1}(t) + \cdots + A_1\lambda(t)\| + \|tA_0 - (1-t)D\| \\ &\leqq \|A_{k-1}\|r^{k-1}(t) + \cdots + \|A_1\|r(t) + s \end{aligned}$$

since $\|x\| = 1$ and $t < 1$. The solution of this polynomial inequality is obviously bounded. The proof is completed. □

Remarks. 1. It follows from a standard degree argument that each circle of solutions of $P(\lambda)x = 0$ in $\mathbb{C}^{n+1}$ with $\|x\| = 1$ is a limit set of one of the component $C(y_{ij})$.

2. If A_k is nonsingular, the proof of Theorem 3.1 can be easily extended such that $H^{-1}(0)$ is uniformly bounded for $t \in [0, 1]$. If A_k is singular, then some components $C(y_{ij})$ will become unbounded as $t \to 1$. This simply indicates that the λ-matrix problem (1.2) does not have nk eigenvalues.

3. For real symmetric eigenvalue problems $Ax - \lambda x = 0$, the homotopy method can be carried out in real arithmetic. In this case, the zero set $H^{-1}(0)$ consists of smooth curves only. Rhee [11] has shown that the local conditioning of the homotopy curves at $t \in (0, 1)$ is affected by two factors: the separation of eigenvalues of the matrix $D + t(A - D)$ and the closeness of D to A. It is interesting to note that a checking criterion can easily be set up to prevent the ODE solver from jumping from one curve to another in the continuation process.

4. Computations. Theorem 3.1 asserts the existence of nk cylindrical tubes $C(y_{ij})$ starting from the hyperplane $t \equiv 0$. We now show how to extract a path from a tube with the intention that this path could be followed numerically and would lead from $t = 0$ to $t = 1$. According to the proof of Theorem 3.1, assertion 3, we could further require that this path be parameterized by the variable t.

Among the many possible ways to define such a path, we choose to consider the following approach.

Let the homotopy function H be a mapping from $\mathbb{R}^{2n} \times \mathbb{R}^2 \times [0, 1)$ to $\mathbb{R}^{2n} \times \mathbb{R}$ so that (x, λ) is identified as a vector in $\mathbb{R}^{2n} \times \mathbb{R}$ and ix a vector in $\mathbb{R}^{2n}$. We define vector fields $(\dot{x}, \dot{\lambda}, 1)$ on $H^{-1}(0)$ by requiring

$$M(x,\lambda,t)\begin{bmatrix}\dot{x}\\ \dot{\lambda}\end{bmatrix} = \begin{bmatrix}(Q(\lambda)-P(\lambda))x\\ 0\end{bmatrix}, \tag{4.1}$$

$$[ix^T, 0]\begin{bmatrix}\dot{x}\\ \dot{\lambda}\end{bmatrix} = 0 \tag{4.2}$$

where $M \in \mathbb{R}^{(2n+1)\times(2n+2)}$ is the corresponding real matrix, defined as in (2.4), of the matrix Q in (3.6). Note that (4.1) is a necessary condition for the vector field $(\dot{x}, \dot{\lambda}, 1)$ to be tangent to the surface $H^{-1}(0)$. Equation (4.2) simply means that the vector field is always perpendicular to the circle of the intersection of the hyperplane $t \equiv$ constant and the tube.

The $(2n + 2) \times (2n + 2)$ real matrix

$$\begin{bmatrix}M(x,\lambda,t)\\ ix^T, 0\end{bmatrix} = \begin{bmatrix}\hat{Q}(x,\lambda,t)\\ a_1, b_1, \cdots, a_n, b_n, 0, 0\\ -b_1, a_1, \cdots, -b_n, a_n, 0, 0\end{bmatrix}$$

is precisely the real representation of the $(n + 1) \times (n + 1)$ complex matrix

$$\begin{bmatrix}R(\lambda,t,D), & R_\lambda(\lambda,t)x\\ x^*, & 0\end{bmatrix}.$$

Therefore, the remaining numerical work is to follow the initial value problem in $\mathbb{C}^n \times \mathbb{C}$:

$$\begin{bmatrix}R(\lambda,t,D), & R_\lambda(\lambda,t)x\\ x^*, & 0\end{bmatrix}\begin{bmatrix}dx/dt\\ d\lambda/dt\end{bmatrix} = \begin{bmatrix}(Q(\lambda)-P(\lambda))x\\ 0\end{bmatrix}, \tag{4.3}$$

$$x(0) = e_i, \quad \lambda(0) = \lambda_{ij}$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, k$.

Remark. Note that the nk homotopy curves we derived are integral curves of the same differential equations subject to different initial values. Since these curves are independent of each other, it is suitable to follow several curves simultaneously on a multiprocessor. Note also that the homotopy function (3.5) does not cause any destruction in the matrix structure of $P(\lambda)$. Combined with a sparse matrix technique, the homotopy method might therefore become attractive for solving large-scale λ-matrix problems.

Remark. In practical computation, it may not be necessary to follow the entire homotopy curve $(x(t), \lambda(t))$ as in (4.3). For example, Rhee [11] proposes an algorithm for real symmetric eigenvalue problems that traces the one-dimensional eigenvalue curves only, whereas the eigenvectors are estimated locally by the Rayleigh quotient iteration. Test results seem to indicate that the overall complexity of the homotopy method for finding all n eigenpairs would be $O(n^2)$ as opposed to $O(n^{2.6})$ of the standard subroutine IMTQL2 in EISPACK. We should point out also that EISPACK is not designed for large-scale matrices, where it can be shown [11] that the conditioning of the eigenvalue curves in the homotopy method is independent of the size of the matrix. With careful coding the homotopy method might be a serious alternative for solving large-scale eigenvalue problems.

We coded the homotopy method (4.3) into an IBM 3081 uniprocessor simply to examine the various behavior of the paths near $t = 1$. With no intention of making this code efficient, we integrated the initial value problem (4.3) by using the subroutine DGEAR found in IMSL. The scalar c and the vector $(d_1, \cdots, d_n)$ were randomly generated. No matrix structure was taken into consideration. The linear equation solver LEQ2C was used to find the vector field in (4.3). The following examples represent a collection of problems we used to experiment with our homotopy method.

Example 1. The following is a symmetric, definite quadratic problem with distinct eigenvalues:

$$P(\lambda) = \begin{bmatrix} -10\lambda^2+\lambda+10, & & & & \\ 2\lambda^2+2\lambda+2, & -11\lambda^2+\lambda+9, & & & \\ -\lambda^2+\lambda-1, & 2\lambda^2+2\lambda+3, & & & \\ \lambda^2+2\lambda+2, & -2\lambda^2+\lambda-1, & & & \\ 3\lambda^2+\lambda-2, & \lambda^2+3\lambda-2, & & & \\ & & \text{(symmetry)} & & \\ & -12\lambda^2+10, & & & \\ & -\lambda^2-2\lambda+2, & -10\lambda^2+2\lambda+12, & & \\ & \lambda^2-2\lambda-1, & 2\lambda^2+3\lambda+1, & -11\lambda^2+3\lambda+10 & \end{bmatrix}.$$

The code found all 10 eigenpairs without any difficulty. The eigenvalues (to seven digits) are: $\{-0.5117619, 0.8799272, 1.465467, -0.7790945, 0.5024152, -1.077167, 0.9365506, -1.004838, 1.956883, -1.271885\}$.

Example 2. The following is a symmetric, definite quadratic problem with multiple eigenvalues:

$$P(\lambda) = \begin{bmatrix} -\lambda^2-3\lambda+1, & & & \\ \lambda^2-1, & -2\lambda^2-3\lambda+5, & \text{(symmetry)} & \\ -\lambda^2-3\lambda+1, & \lambda^2-1, & -2\lambda^2-5\lambda+2, & \\ -2\lambda^2-6\lambda+2, & 2\lambda^2-2, & -4\lambda, & -9\lambda^2-19\lambda+14 \end{bmatrix}.$$

This problem has double eigenvalues 1 and -2, and simple eigenvalues $-4 \pm \sqrt{19}$ and $-4 \pm \sqrt{18}$. The code had no difficulty in accurately locating all eigenpairs (although it took a slightly extra effort to tackle the multiple eigenvalue cases). The computed eigenvectors associated with multiple eigenvalues were linearly independent.

Example 3. The following is an unsymmetrical cubic problem with singular leading coefficients:

$$P(\lambda) = \begin{bmatrix} 5\lambda^3 + \lambda + 1 & \lambda^3 + 1, & \lambda^3 + 1 \\ \lambda^3 + \lambda, & 5\lambda^3 + \lambda + 1, & \lambda^3 + 1 \\ \lambda^3 + \lambda, & 5\lambda^3 + \lambda, & \lambda^3 + \lambda + 1 \end{bmatrix}.$$

The problem actually has only seven eigenvalues since $\det(P(\lambda)) = 27\lambda^7 + 4\lambda^6 + 9\lambda^5 + 9\lambda^4 + 6\lambda^3 + \lambda^2 + \lambda + 1$. The code had no problem in locating these seven eigenpairs. The eigenvalues (to six digits) are $(0.307991 \pm 0.686745i, -0.453629 \pm 0.460837i, 0.327145 \pm 0.474411i, -0.529683\}$. Two of the nine homotopy curves escaped to infinity (with $\|x\| = 1$ always) as t approaches 1. This deceived the code into giving two large extraneous eigenvalues and their associated eigenvectors.

Example 4. The following is a quadratic problem with high multiplicity of eigenvalues and high degeneracy of eigenvectors:

$$P(\lambda) = \begin{bmatrix} (\lambda-1)(\lambda-4), & 5-2\lambda, & 0 \\ 0, & (\lambda-1)(\lambda-4), & 5-2\lambda \\ 0, & 0, & (\lambda-1)(\lambda-4) \end{bmatrix}.$$

Obviously the eigenvalues of this problem are 1 and 4 only, and each eigenvalue is of multiplicity 3. Furthermore, this problem has only one eigenvector. With local tolerance TOL = 10^{-6} in DGEAR, the code was returned with all six curves being convergent. However, the accuracy was only about 10^{-2}. This was due to a high-order bifurcation occurring at $t = 1$.

REFERENCES

[1] M. T. CHU, *A simple application of the homotopy method to symmetric eigenvalue problems*, Linear Algebra Appl., 59 (1984), pp. 85–90.

[2] ———, *A note on the homotopy method for linear algebraic eigenvalues problems*, Linear Algebra Appl., to appear.

[3] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[4] P. LANCASTER, *Lambda-Matrices and Vibrating Systems*, Pergamon Press, Oxford, 1966.

[5] ———, *A review of numerical methods for eigenvalue problems nonlinear in parameter*, Internat. Ser. Numer. Math., 38 (1977), pp. 43–67.

[6] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Second edition, Academic Press, Orlando, FL, 1985.

[7] T. Y. LI AND T. SAUER, *Homotopy method for generalized eigenvalue problems*, Linear Algebra Appl., 91 (1987), pp. 65–74.

[8] T. Y. LI, T. SAUER, AND J. A. YORKE, *Numerical solution of a class of deficient polynomial systems*, SIAM J. Numer. Anal., 24 (1987), pp. 435–451.

[9] J. MILNOR, *Singular Points of Complex Hypersurfaces*, Ann. of Math. Stud., 61, Princeton University Press, Princeton, NJ, 1968.

[10] G. PETERS AND J. H. WILKINSON, *$Ax = \lambda Bx$ and the generalized eigenproblem*, SIAM J. Numer. Anal., 7 (1970), pp. 479–492.

[11] N. H. RHEE, *The homotopy method for the symmetric eigenvalue problems*, Ph.D. thesis, Michigan State University, East Lansing, MI, 1987.

[12] A. RUHE, *Algorithms for the nonlinear eigenvalue problem*, SIAM J. Numer. Anal., 4 (1973), pp. 674–689.

[13] D. D. SCOTT, *Solving sparse symmetric definite quadratic matrix problems*, BIT, 21 (1981), pp. 475–480.

[14] B. L. VAN DER WAERDEN, *Modern Algebra*, Vol. 1, Frederick Ungar, New York, 1949.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 4, October 1988

007

CYCLE LENGTHS IN A^kb*

CHARLES M. GRINSTEAD†

Abstract. Let A be a nonnegative, $n \times n$ matrix, and let b be a nonnegative, $n \times n$ vector. Let S be the sequence $\{A^kb\}$, $k = 0, 1, 2, \cdots$. Define $m(A, b)$ to be the length of the cycle of zero-nonzero patterns into which S eventually falls. Define $m(A)$ to be the maximum, over all nonnegative b of $m(A, b)$. Finally, define $m(n)$ to be the maximum, over all nonnegative, $n \times n$ matrices A of $m(A)$. This paper shows given A and b, that $m(A, b)$ is a divisor of a certain number, which is determined by the structure of A and b. It is also shown that $\log m(n) \sim (n \log n)^{1/2}$.

Key words. positive matrices, symmetric group, Prime Number Theorem

AMS(MOS) subject classifications. 15A48, 10H25

Let A be a nonnegative, $n \times n$ matrix, and let b be a nonnegative, $n \times 1$ vector. In this paper, we are concerned with the zero-nonzero patterns in the sequence $S = \{A^kb\}$, $k = 0, 1, 2, \cdots$. Since there are 2^n possible patterns for an $n \times 1$ vector, the sequence S must eventually fall into a cycle, and the length of the cycle is at most 2^n. We define $m(A, b)$ to be the length of this cycle. We also define $m(A)$ to be the maximum, over all b of $m(A, b)$. Finally, we define $m(n)$ to be the maximum, over all nonnegative, $n \times n$ matrices A of $m(A)$. Given A and b, we will show that $m(A, b)$ is a divisor of a certain number, which is determined by the structure of A and b. We will also show that

$$\log (m(n)) \sim \sqrt{(n \log n)}.$$

Each nonnegative, $n \times n$ matrix A corresponds to a directed graph $G(A)$ with vertices $1, 2, \cdots, n$, and with an edge from j to i if $a_{ij} > 0$. (If $a_{ii} > 0$, then there is a loop at vertex i.) This is *not* the usual definition. However, the present definition aids in the exposition. We note that our graph could be obtained by applying the usual definition to the transpose of A. Each nonnegative vector b corresponds to a subset $P(b)$ of vertices, defined by $i \in P(b)$ if and only if $b_i > 0$. Given a vector b, the set corresponding to Ab is the set of all vertices of distance one from $P(b)$, i.e., all vertices j such that there is a vertex $i \in P(b)$ and an edge (i, j) in $G(A)$. If we define

$$\Gamma^{(k)}(v) = \{w: \text{there is a path of length } k \text{ from } v \text{ to } w\},$$
$$\Gamma^{(k)}(T) = \bigcup_{v \in T} \Gamma^{(k)}(v),$$

then we have

$$P(A^kb) = \Gamma^{(k)}(P(b)).$$

A subgraph H of a graph G is said to be *strongly connected* if, for any two (not necessarily distinct) vertices in H, there is a path in H from each vertex to the other. A subgraph is a *strongly connected component* (scc) if it is a maximal strongly connected subgraph of G. It is easy to show that the scc's are pairwise disjoint. They need not, however, partition the graph.

In terms of matrices, given A, we let P be a permutation matrix such that PAP^T is in block lower triangular form, with square matrices on the diagonal. If no such P exists,

* Received by the editors June 9, 1986; accepted for publication (in revised form) March 29, 1988. This research was partly supported by National Science Foundation grant DMS-8406451.

† Swarthmore College, Swarthmore, Pennsylvania 19081.

other than $P = I$, then the matrix A is said to be irreducible. If we find P such that each diagonal block is irreducible, then the diagonal blocks of size greater than 1×1, together with the 1×1 nonzero diagonal blocks, correspond to the scc's of $G(A)$.

LEMMA 1. *Let $S_1, S_2, \cdots$ be a sequence of subsets of a finite set S, and let d be a positive integer. Suppose that for all j and for all sufficiently large h, we have $S_j \subset S_{j+hd}$. Then the sequence has an eventual cycle whose length divides d.*

Proof. Let j be any nonnegative integer less than d. Let S^j be the set of all $v \in S$ such that $v \in S_{j+hd}$ for some $h \geqq 0$. For each $v \in S^j$, there exists an integer h_v such that $v \in S_{j+hd}$ for all $h \geqq h_v$. Let h_0 be the maximum of the h_v, taken over all $v \in S^j$. Then $S_{j+hd} = S^j$ for all $h \geqq h_0$. Thus, the sequence has an eventual cycle consisting of S^0, $S^1, \cdots, S^{(d-1)}$. This implies that the sequence has an eventual cycle the length of which divides d. □

We define the *index* of a graph to be the greatest common divisor of the lengths of the circuits in the graph. The index of a nonnegative matrix A is then defined to be the index of $G(A)$. (We note that this is not the usual definition of the index of a matrix.) Let b be a nonnegative vector, and for $j \geqq 0$ denote $P(A^j b)$ by P_j. We first examine $m(A, b)$ in the case that A is irreducible.

LEMMA 2. *Let A be a nonnegative, irreducible $n \times n$ matrix with index d. Then, for all b, $m(A, b)$ divides d.*

Proof. It is well known (see [5, Thm. 2.9, p. 49]) that the greatest common divisor of the lengths of the circuits through any vertex v is d, independent of the vertex v. It is also well known that if we call these circuit lengths $c_1, c_2, \cdots, c_j$, then there is a multiple of d, say $N_v d$, such that every multiple of d greater than or equal to $N_v d$ can be written as a nonnegative linear combination of the $\{c_i\}$. If we let $N = \max_{v \in G} N_v$, then every vertex in G is on a circuit of length hd, for all $h \geqq N$. Thus, if $v \in P_j$, then $v \in P_{j+hd}$ for all sufficiently large h. Then Lemma 1 applies. This completes the proof. □

LEMMA 3. *Let A be a nonnegative, $n \times n$ matrix, and let C be an* scc *in $G(A)$. Let b be a nonnegative vector, and, as before, let $P_j = P(A^j b)$ and $P_0 = P(b)$. Let the index of C be d. Then the sequence $\{P_j \cap C\}$ eventually repeats with a cycle length that divides d.*

Proof. Let j be a positive integer, and let v be any vertex in $P_j \cap C$. Since C is an scc with index d, v is on a circuit in C of length hd, for all sufficiently large h. This means that v is in $P_{j+hd} \cap C$ for all sufficiently large h. Thus, from Lemma 1, we see that the sequence $\{P_j \cap C\}$ eventually repeats with a cycle length that divides d. □

THEOREM 1. *Let A be a nonnegative, $n \times n$ matrix, and let $G(A)$ have* scc's *$C_1, C_2, \cdots, C_k$. Let b be any nonnegative vector. Suppose that the sequence $\{P_j \cap C_i\}$ has an eventual cycle of length r_i. Then $m(A, b)$ equals the least common multiple of the r_i.*

Proof. Let r equal the least common multiple (lcm) of the r_i. First we show that $m(A, b)$ divides r. To show this, we shall show that for all vertices v in $G(A)$, the sequence $\{P_j \cap \{v\}\}$ has an eventual cycle whose length divides r.

Let $v \in C_i$ for some i. If $v \in P_k$ for some $k \geqq 0$, then for some positive integer j and for all sufficiently large h, we have $v \in P_{j+hr}$, since r is a multiple of r_i. So Lemma 1 implies that the sequence $\{P_j \cap \{v\}\}$ has an eventual cycle whose length divides r. If $v \neg\in P_k$ for all $k \geqq 0$, then the sequence $\{P_j \cap \{v\}\}$ has a cycle of length 1.

Now suppose that v is not an element of any C_i. Let $C_{i1}, C_{i2}, \cdots, C_{ik}$ be the scc's containing vertices that have paths to v. If $k = 0$, then v is in no P_i with $i \geqq n$. Finally, if $k > 0$, then for $j \geqq n$, $v \in P_j$ if and only if there is a v^* in $C_{is} \cap P_t$, for some $s \leqq k$, and a path of length $(j - t)$ from v^* to v. (We need to assume that $j \geqq n$ because, for smaller values of j, the fact that $v \in P_j$ could be due to v being at the end of a path of length j from a vertex in P_0 not in any C_i.) Since $v^* \in C_{is}$, the eventual cycle in the

sequence $\{P_j \cap \{v^*\}\}$ has a length that divides r_{is}. This means that the contribution of v^* to the eventual cycle in the sequence $\{P_j \cap \{v\}\}$ has a length that divides r_{is}. Since the contributions of all other vertices in $C_{i1}, C_{i2}, \cdots, C_{ik}$ to the sequence $\{P_j \cap \{v\}\}$ are similar, we see that the sequence $\{P_j \cap \{v\}\}$ has an eventual cycle whose length certainly divides r. So we have shown that for all vertices in $G(A)$, either they appear in none of the P_j for sufficiently large j, or they appear with a pattern having a length that divides r. This means that $m(A, b)$ divides r.

Since the eventual cycle of the sequence $\{P_j \cap C_i\}$ is of length r_i, it is clear that r_i must divide $m(A, b)$. □

COROLLARY 1. *Let A be a nonnegative, $n \times n$ matrix, and let $G(A)$ have* scc's *$C_1, C_2, \cdots, C_l$ with indices $d_1, d_2, \cdots, d_l$, respectively. Let b be any nonnegative vector. Let T be the set of* scc's *that intersect at least one element of the sequence $\{P_k\}$. Let d_T equal the least common multiple of the set of d_i's corresponding to the C_i's in T. Then $m(A, b)$ divides d_T.*

Proof. If $C_i \in T$, and we let r_i equal the length of the eventual cycle of the sequence $\{P_j \cap C_i\}$, then using Lemma 3, we know that r_i divides d_i. If $C_i \neg\in T$, then the length of the eventual cycle of the sequence $\{P_j \cap C_i\}$ is 1. Thus, since $m(A, b)$ equals the lcm of the r_i's corresponding to the C_i's in T, we see that $m(A, b)$ divides d_T. □

We now turn our attention to the maximization of $m(A)$ over all nonnegative $n \times n$ matrices A. Given positive integers $d_1, d_2, \cdots, d_k$, with sum n, it is possible to construct a nonnegative, $n \times n$ matrix A and a nonnegative vector b such that $m(A, b) = \text{lcm}\ (d_1, d_2, \cdots, d_k)$. We accomplish this by letting $G(A)$ be the disjoint union of circuits of lengths $d_1, d_2, \cdots, d_k$, and letting P_0 be a set containing exactly one vertex from each circuit.

Next, we note that $d_i \leqq |C_i|$ for each i, and that

$$\sum_{i=1}^{k} |C_i| \leqq n,$$

so we know that

$$\sum_{i=1}^{k} d_i \leqq n.$$

Thus, we wish to maximize the lcm of $d_1, d_2, \cdots, d_k$ over all sets of positive integers with sum not exceeding n. The maximum value will be $m(n)$. Let us call a set $\{d_i\}$ whose lcm is this maximum value an n-extremal set. We note that this problem can be stated as follows. Among all elements of the symmetric group S_n, which elements have the largest order, and what is their order? In terms of the original problem, the group elements of largest order are those that, when written as a product of disjoint cycles, have cycle lengths forming an n-extremal set. The order of these elements is $m(n)$. This problem was studied by Landau (see [3, Vol. 1, pp. 222–229]), who proved Theorem 2 below (see also [4]). We will give a shorter proof of this result.

LEMMA 4. *For every $n \geqq 1$, there exists an n-extremal set X such that each element of X is either a prime power or the number* 1.

Proof. Assume that X is an n-extremal set containing an integer r, which is neither 1 nor a prime power. Then r is divisible by at least two different primes, p and q. Suppose that the powers of p and q appearing in the prime factorization of r are p^y and q^z. In X, if we replace r by the integers p^y, q^z, and $(r/p^y q^z)$, then the lcm remains unchanged, and the sum of the elements of X decreases by the quantity

$$\Delta = r - \left(p^y + q^z + \left(\frac{r}{p^y q^z}\right)\right).$$

It remains to show that $\Delta \geqq 0$. We have

$$\Delta = r\left(1 - \frac{1}{p^y q_z}\right) - p^y - q^z$$

$$\geqq r\frac{5}{6} - p^y - q^z.$$

Since p^y and q^z divide r,

$$\Delta \geqq r\frac{5}{6} - \frac{r}{p^y} - \frac{r}{q^z}$$

$$= r\left(\frac{5}{6} - \frac{1}{p^y} - \frac{1}{q^z}\right)$$

$$\geqq r\left(\frac{5}{6} - \frac{1}{2} - \frac{1}{3}\right)$$

$$= 0.$$

□

Let p_i denote the ith prime. At first glance, it might seem that an n-extremal set should consist of $p_1, p_2, \cdots, p_k$, where k is the largest prime such that the sum of the first k primes does not exceed n. However, it is easy to show that this is not, in general, the best way to proceed. As an example, suppose that n is the sum of all of the primes not exceeding the prime 1231. The numbers 2, 3, 5, and 1231 have the same sum as the numbers 2^9 and 3^6, but the lcm of the second set is larger than the lcm of the first set. So, by replacing the first set with the second set, we obtain a set with a larger lcm.

It is nevertheless the case that by taking the first k primes, we obtain a set whose lcm has, asymptotically, the same logarithm as $m(n)$.

THEOREM 2. *Given a positive integer n, let k be the largest integer such that*

$$\sum_{i=1}^{k} p_i \leqq n.$$

Then, $\log(m(n)) \sim \sum_{i=1}^{k} \log(p_i)$. *Furthermore,* $k \sim 2\sqrt{(n)}/\sqrt{(\log(n))}$, *so*

$$\log(m(n)) \sim (\sqrt{n})(\sqrt{(\log(n))}).$$

Before proving this theorem, we need the following lemma.

LEMMA 5. *Let T and T' be two sets of real numbers with the following properties:*

(i) *Each element of T is less than each element of T'.*

(ii) *Every element of both sets is at least as large as e.*

(iii) *The sum of the elements in T is at least as large as the sum of the elements of T'.*

Then the product of the elements of T is at least as large as the product of the elements of T'.

Proof. Let B be a real number at least as large as each element of T and less than each element of T'. We note that B can be chosen to be at least e. Let S and S' be the sums of the elements in the sets T and T', respectively. Let P and P' be the products of the elements in the sets T and T', respectively.

First, fix $|T'| = k$. If two elements of T' are unequal, we can make P' larger without affecting S', by replacing each of the two elements by their average. Thus, we may assume that all of the elements in T' are equal. Then their common value is (S'/k),

and $P' = (S'/k)^k$. Since each element in T' is greater than B, it is easy to show that $P' \leqq B^{(S'/B)}$.

If q is any real number such that $e \leqq q \leqq B$, then it is easy to check that $q \geqq B^{(q/B)}$. Thus, if the elements of T are $q_1, q_2, \cdots, q_k$, then the product of the elements of T is at least $(B^{(q_1/B)})(B^{(q_2/B)}) \cdots (B^{(q_k/B)})$, which equals $B^{(S/B)}$. Since $S \geqq S'$, we have $P \geqq P'$, which completes the proof. □

Proof of Theorem 2. The Prime Number Theorem implies that $p_i \sim i \log(i)$ (see [2, p. 10]). Using this, it is easy to show that

$$\sum_{i=1}^{k} p_i \sim \left(\frac{1}{2}\right) k^2 \log(k).$$

Thus, we have $n \sim (\frac{1}{2})k^2 \log(k)$, which implies that $\log(k) \sim (\frac{1}{2}) \log(n)$. Hence,

$$k \sim 2(\sqrt{(n)})/(\sqrt{(\log(n))}).$$

We note that this implies that

$$p_k \sim (\sqrt{(n)})(\sqrt{(\log(n))}).$$

Now assume for the moment that n is the sum of the first k primes. Let S be the set of primes not exceeding p_k, and let S' be an n-extremal set. The sum of the elements in S' is then less than or equal to the sum of the elements in S, which equals n. Let T' be the set of all elements of S' which are powers of primes p_j such that $j > k$. Let T be the set of all primes p_i in S such that no power of p_i appears in S'. Since each prime in $S - T$ appears to the first power in $S - T$, and appears to at least the first power in $S' - T'$, and since the sum of the elements in S is at least as great as the sum of the elements in S', we must have that

$$\sum_{q_j \in T'} q_j \leqq \sum_{p_i \in T} p_i,$$

where each q_j in the left-hand sum represents a prime power. We further note that each q_j in T' is greater than p_k, and that each p_i in T is less than or equal to p_k. We now note that Lemma 5 applies, except that one of the p_i in T might be the prime 2. If we temporarily change it to a 3, then, using Lemma 5, we see that the product of the elements of T is at least as great as the product of the elements in T'. Changing the 3 back to a 2 certainly does not affect the dominant term in the estimation for the logarithm of the product of the elements in S'. Thus in S', if the elements in T' are replaced by the elements in T, the product of the elements of the new S' is at least two-thirds as large as the product of the elements in the old S', hence we may assume that S' contains no powers of any prime greater than p_k. At this point we emphasize that S' may have a sum that exceeds n. Nevertheless, each element in S' is less than or equal to n. Using this fact, we shall show that the logarithm of the product of the elements in S' does not exceed $\sqrt{(n)}(\log n)$. We estimate:

$$\begin{aligned}
\log\left(\prod_{q_i \in S'} q_i\right) &= \log\left(\prod_{p_i \leqq \sqrt{n}} p_i^{\alpha_i}\right) + \log\left(\prod_{\substack{p_i > \sqrt{n} \\ i \leqq k}} p_i\right) \\
&\leqq \log(n^{\pi(\sqrt{n})}) + \sum_{\substack{p_i > \sqrt{n} \\ i \leqq k}} \log(p_i) \\
&\sim \frac{\sqrt{n}(\log n)}{\log \sqrt{n}} + \sqrt{n}\sqrt{\log n} - \sqrt{n} \\
&\sim \sqrt{n}\sqrt{\log n}.
\end{aligned}$$

We now show that this bound is achieved by the elements in S. We have

$$\sum_{i=1}^{k} \log(p_i) \sim p_k$$

(see [2, Thms. 420, 434]). Also, since $n = \sum_{i=1}^{k} p_i$ and $p_i \sim i(\log i)$, it is easy to show that

$$\begin{aligned} p_k &\sim (2\sqrt{(n)}/\sqrt{(\log n)})\log(2\sqrt{(n)}/\sqrt{(\log n)}) \\ &\sim \sqrt{n}\sqrt{\log n}. \end{aligned}$$

Thus, if n is the sum of the first k prime numbers, then

$$\log(m(n)) \sim \sqrt{n}\sqrt{\log n}.$$

Finally, we relax the assumption on n. Assume instead that

$$n_1 = \sum_{i=1}^{k} p_i < n \leqq \sum_{i=1}^{k+1} p_i = n_2.$$

Since $m(n)$ is clearly monotonic in n, we have $m(n_1) \leqq m(n) \leqq m(n_2)$, but it is easy to check that $m(n_1) \sim m(n_2)$, which completes the proof. □

A comment on the conclusion of the theorem is in order. While it would be nicer to obtain a function to which $m(n)$ is asymptotic, it seems unlikely that this can be done. The task would require an estimation similar to the estimation of the product of the first k primes. Let P be this product. We would have to obtain an estimate of the following form:

$$\sum_{i=1}^{k} \log(p_i) = f(k) + o(1),$$

for then we would be able to say that $P \sim e^{f(k)}$. Although it is known that the above sum is asymptotic to p_k, at this time we cannot even say that the error term is $O(p_k^{\delta})$ for even one value of $\delta < 1$.

Acknowledgement. The author thanks the referees for their numerous helpful observations.

REFERENCES

[1] P. G. COXSON AND L. LARSON, *Monomial patterns in the sequence $A^k b$*, preprint.
[2] G. H. HARDY AND E. M. WRIGHT, *An Introduction to the Theory of Numbers*, Clarendon Press, Oxford, 1960.
[3] E. LANDAU, *Handbuch der Lehre von der Verteilung der Primzahlen*, *Teubner*, Leipzig, Stuttgart, 1909.
[4] W. MILLER, *The maximum order of an element in a finite symmetric group*, Amer. Math. Monthly, 94 (1987), pp. 497–506.
[5] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 4, October 1988

008

EIGENVALUES AND CONDITION NUMBERS OF RANDOM MATRICES*

ALAN EDELMAN†

Abstract. Given a random matrix, what condition number should be expected? This paper presents a proof that for real or complex $n \times n$ matrices with elements from a standard normal distribution, the expected value of the log of the 2-norm condition number is asymptotic to $\log n$ as $n \to \infty$. In fact, it is roughly $\log n + 1.537$ for real matrices and $\log n + 0.982$ for complex matrices as $n \to \infty$. The paper discusses how the distributions of the condition numbers behave for large n for real or complex and square or rectangular matrices. The exact distributions of the condition numbers of $2 \times n$ matrices are also given.

Intimately related to this problem is the distribution of the eigenvalues of Wishart matrices. This paper studies in depth the largest and smallest eigenvalues, giving exact distributions in some cases. It also describes the behavior of all the eigenvalues, giving an exact formula for the expected characteristic polynomial.

Key words. characteristic polynomial, condition number, eigenvalues, random matrices, singular values, Wishart distribution

AMS(MOS) subject classification. 15A52

1. Introduction.

What is the condition number of a random matrix? Though we were originally motivated by this question, the problem quickly becomes one of studying the eigenvalues of a related random matrix.

This application of random eigenvalues originally appeared in a classic paper by von Neumann and Goldstine [22]. Further applications can be found in statistics and physics (see, e.g., [7], [25]). Statisticians use random eigenvalues in principal component analysis, multiple discriminant analysis, and canonical correlation analysis. Physicists model nuclear levels with eigenvalues.

When speaking of a random matrix, we will focus on the Gaussian and Wishart distributions. We say that a matrix X has the Gaussian distribution if each element of the matrix comes from an independent standard normal distribution. We obtain Wishart matrices from Gaussian matrices by forming XX^T. Wishart matrices are of intrinsic interest because they are essentially the sample covariance matrices for multivariate Gaussian distributions, as discussed in books on multivariate statistics such as [25].

Various researchers have investigated the eigenvalues of a Wishart matrix from a number of points of view. If we take a large matrix from a Wishart distribution, we may sort and plot the eigenvalues against their position index. A theory of what the picture should be is developed in [13], [16], [21], and [23]. Estimates of the largest and smallest eigenvalues are given in [9] and [17]. A complicated expression for the distribution of the largest eigenvalue is given in [19] and for the smallest eigenvalue in [15].

Our question about condition numbers was introduced in a precise format in [18]. In effect, Smale asks for the expected geometric mean of the condition number of a Gaussian matrix. Precisely, let X_n be an $n \times n$ matrix whose elements are independent standard normal random variables. Let $\kappa_{X_n} = \|X_n\| \|X_n^{-1}\|$ be its condition number in the 2-norm. What is the expected value of $\log \kappa_{X_n}$? The reason we use $\log \kappa_{X_n}$ is that this quantity is the measure of the loss of numerical precision (see [6]). The result of directly averaging the condition number, on the other hand, is known to be infinite. Kostlan and

* Received by the editors November 25, 1987; accepted for publication (in revised form) April 4, 1988. This research was supported by a Hertz Foundation Fellowship and by an IBM Faculty Development Award to Lloyd N. Trefethen.

† Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

Ocneanu (see [18]) obtained some estimates showing that for all $\varepsilon > 0$, when n is sufficiently large,

$$\frac{2}{3} - \varepsilon \leqq \frac{E(\log \kappa_{X_n})}{\log n} \leqq \frac{5}{2} + \varepsilon.$$

Kostlan has communicated to me a new result that raises the lower bound to 1 [14]. In the present paper, we show that this new result is sharp: $E(\log \kappa_{X_n}) \sim \log n$ as $n \to \infty$. The same leading behavior holds for complex matrices, but we have more precise estimates. We also explore asymptotic results for rectangular matrices.

A natural first step in conducting this investigation was to run some numerical experiments. In Table 1.1, we list the result of averaging log condition numbers from random samples of 1000 square matrices of dimension equal to various powers of 2. Also listed are the results for 1000 matrices of dimensions 100×200. The data for square matrices clearly suggest $E(\log \kappa_{X_n}) \sim \log n$ for both the real and complex cases, and we might perhaps predict that $E(\log \kappa_{X_n}) = \log n + c + o(1)$ for some constant c. In § 6, we derive the constant c (≈ 1.537 for real matrices and ≈ 0.982 for complex matrices). We also show that for large (real or complex) matrices the condition number depends on the ratio of rows to columns m/n. For example, matrices with twice as many columns as rows have an expected log condition number asymptotic to 1.76. It is of interest that this value is finite. In the table we see that the asymptotic result gives a usable approximation for the finite case.

In Table 1.2, we summarize our results about condition numbers in the limit $n \to \infty$. (Please consult the text for details not explained here.) The values listed are the exponentials of the expected logarithms of three random variables: the condition number of the Gaussian matrix and the largest and smallest eigenvalues of the related Wishart matrix. Note that this first quantity is the ratio of the square root of the other two quantities. As a kind of table of contents, the table lists where these results are stated explicitly in the text. K_2 is in fact $2e^{\gamma/2}$, where γ is Euler's constant, ≈ 0.5772. K_1 is a little more complicated. It is given by γ and a readily evaluated definite integral. For the rectangular matrices, the variable y denotes the ratio m/n, where $0 < y < 1$.

For the special case of real and complex $2 \times n$ matrices we can specify exactly the distributions of condition numbers and eigenvalues; these results are reported in § 7. We comment about the tail of the condition number distribution in § 8. We look at the complete spectrum of a Wishart matrix in § 9 and derive further exact distributions in § 10.

TABLE 1.1
Average log condition numbers.

	Real		Complex	
n	Avg.	Avg. $-$ log n	Avg.	Avg. $-$ log n
2	1.53	0.84	1.19	0.49
4	2.63	1.25	2.09	0.70
8	3.46	1.38	2.91	0.83
16	4.24	1.47	3.65	0.88
32	4.93	1.47	4.35	0.88
64	5.64	1.48	5.06	0.90
128	6.44	1.59	5.78	0.93
256	7.04	1.49	6.50	0.96
100×200	1.72		1.67	

TABLE 1.2
Exponentials of expected logs ($K_1 \approx 4.65$, $K_2 \approx 2.67$).

		Real	Complex
Square	κ	$K_1 n$ Thm. 6.1	$K_2 n$ Thm. 6.2
	$\lambda_{\max}$	$4n$ Prop. 4.1	$8n$ Prop. 4.2
	$\lambda_{\min}$	$\frac{4}{K_1^2 n}$ Cor. 3.2	$\frac{8}{K_2^2 n}$ Cor. 3.4
Rectangular	κ	$\frac{1+\sqrt{y}}{1-\sqrt{y}}$ Thm. 6.3	$\frac{1+\sqrt{y}}{1-\sqrt{y}}$ Thm. 6.3
	$\lambda_{\max}$	$n(1+\sqrt{y})^2$ Prop. 4.1	$2n(1+\sqrt{y})^2$ Prop. 4.2
	$\lambda_{\min}$	$n(1-\sqrt{y})^2$ Prop. 5.1	$2n(1-\sqrt{y})^2$ Prop. 5.2

2. Gaussian and Wishart matrices. We are interested in rectangular Gaussian matrices, that is, $m \times n$ matrices all of whose components are independent standard normal variables. We denote such a random matrix (or its distribution) by $G(m, n)$. $G(m, n)$ has the symmetry property that it is invariant under orthogonal transformations (i.e., isotropic).

A derived random matrix is the $m \times m$ Wishart matrix $W(m, n)$ defined by $M = XX^T$, where X has the distribution $G(m, n)$. We will focus on the eigenvalues of M, $\lambda_{\max} = \lambda_1 \geqq \cdots \geqq \lambda_m = \lambda_{\min} \geqq 0$, since they are the squares of the singular values of X, and the 2-norm condition number of X is $\sqrt{\lambda_{\max}/\lambda_{\min}}$.

Remarkably enough, the exact joint density function for the m eigenvalues of M can be written as

$$K_{n,m} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}\lambda_i\right)\prod_{i=1}^{m}\lambda_i^{(n-m-1)/2}\prod_{i<j}(\lambda_i-\lambda_j)\,d\lambda_1\cdots d\lambda_m, \tag{1}$$

where

$$K_{n,m}^{-1} = \left(\frac{2^n}{\pi}\right)^{m/2}\prod_{i=1}^{m}\Gamma\left(\frac{n-i+1}{2}\right)\Gamma\left(\frac{m-i+1}{2}\right) \tag{2}$$

(see [12] or [25]).

We may further define complex Wishart matrices $\tilde{M} = \tilde{X}\tilde{X}^T$, where $\tilde{X}$ is of the form $X_1 + iX_2$, with X_1, X_2 each independent and with distribution $G(m, n)$. Let $\tilde{G}(m, n)$ and $\tilde{W}(m, n)$ denote the distributions of $\tilde{X}$ and $\tilde{M}$, respectively. In this case also, the exact joint density function for the m eigenvalues is known [12]:

$$\tilde{K}_{n,m} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}\lambda_i\right)\prod_{i=1}^{m}\lambda_i^{n-m}\prod_{i<j}(\lambda_i-\lambda_j)^2\,d\lambda_1\cdots d\lambda_m, \tag{3}$$

where

$$\tilde{K}_{n,m}^{-1} = 2^{mn}\prod_{i=1}^{m}\Gamma(n-i+1)\Gamma(m-i+1). \tag{4}$$

3. The smallest eigenvalue of $W(n, n)$ and $\tilde{W}(n, n)$. In Theorem 3.1, we show that the probability density function (pdf) for the smallest eigenvalue, $\lambda_{\min}$ of a matrix from $W(n, n)$ is given exactly by a confluent hypergeometric function of a single variable. The exact distributions of the largest and smallest eigenvalues of Wishart matrices are known in certain cases (see [15] and [19]), but these distributions are given as zonal polynomials or hypergeometric functions of matrix arguments that are computationally unwieldy. In contrast, the function described below in Theorem 3.1 is readily calculated numerically by equations 13.1.2 and 13.1.3 in [2].

In Corollary 3.1, we will observe that $n\lambda_{\min}$ converges in distribution to a random variable whose distribution has a simple form. From the limiting distribution, we will analyze the asymptotic behavior of $\log \lambda_{\min}$, which is the key factor in analyzing $E(\log \kappa)$, the expected log condition number.

THEOREM 3.1. *If M_n has the distribution $W(n, n)$, $n \geqq 1$, then the* pdf *of $\lambda_{\min}$ is given by*

$$f_{\lambda_{\min}}(\lambda) = \frac{n}{2^{n-1/2}} \frac{\Gamma(n)}{\Gamma(n/2)} \lambda^{-1/2} e^{-\lambda n/2} U\left(\frac{n-1}{2}, -\frac{1}{2}, \frac{\lambda}{2}\right).$$

When $a > 0$ and $b < 1$, the Tricomi function, $U(a, b, z)$, is the unique solution to Kummer's equation

$$z\frac{d^2w}{dz^2} + (b - z)\frac{dw}{dz} - aw = 0 \tag{5}$$

satisfying $U(a, b, 0) = \Gamma(1 - b)/\Gamma(1 + a - b)$ and $U(a, b, \infty) = 0$.

Proof. Integrating (1), we obtain

$$f_{\lambda_{\min}}(\lambda) = K_n \lambda^{-1/2} e^{-\lambda/2} \int_{R_\lambda} \exp\left(-\sum_{i=1}^{n-1} \frac{\lambda_i}{2}\right) \prod_{1 \leqq i < j \leqq n-1} (\lambda_i - \lambda_j) \prod_{i=1}^{n-1} (\lambda_i - \lambda) \lambda_i^{-1/2} \, d\lambda_i,$$

where $R_\lambda = \{\lambda_1 \geqq \cdots \geqq \lambda_{n-1} \geqq \lambda\} \subseteq R^{n-1}$ and $K_n^{-1} = \pi^{-n/2} 2^{n^2/2} \prod_{i=1}^{n} \Gamma(i/2)^2$.

The first trick is the transformation $x_i = \lambda_i - \lambda$,

$$f_{\lambda_{\min}}(\lambda) = \frac{K_n}{(n-1)!} \lambda^{-1/2} e^{-\lambda n/2} \int_{R_+^{n-1}} \prod_{i=1}^{n-1} (x_i + \lambda)^{-1/2} \Delta \, d\mu(x_1) \cdots d\mu(x_{n-1}),$$

where $\Delta = \prod_{1 \leqq i < j \leqq n-1} |x_i - x_j|$, $d\mu(x) = xe^{-x/2}$, and the integration takes place over $R_+^{n-1} = \{(x_1, \cdots, x_{n-1}) : x_i \geqq 0\}$. Let $w(\lambda)$ denote the integral above. Our goal is to show that w satisfies (5).

Let $\Delta = \delta\Delta_2$, where $\delta = \prod_{i=2}^{n-1} |x_1 - x_i|$ and $\Delta_2 = \prod_{2 \leqq i < j \leqq n-1} |x_i - x_j|$. Further, let $f_j^{a,b} = x_j^a (x_j + \lambda)^b$ and $g_j = \prod_{i=j}^{n-1} (x_i + \lambda)^{-1/2}$. Last, let $d\Omega = d\mu(x_1) \cdots d\mu(x_{n-1})$ and $d\Omega_2 = d\mu(x_2) \cdots d\mu(x_{n-1})$. Below we express w, w', and w'' using this notation. All the integrations are over R_+^{n-1}, and symmetry is used when possible.

$$w = \int g_1 \Delta \, d\Omega,$$

$$w' = -\frac{n-1}{2} \int f_1^{0,-3/2} g_2 \Delta \, d\Omega,$$

$$w'' = \frac{(n-1)(n-2)}{4} \int f_1^{0,-3/2} f_2^{0,-3/2} g_3 \Delta \, d\Omega + \frac{3}{4}(n-1) \int f_1^{0,-5/2} g_2 \Delta \, d\Omega.$$

Since $g_1 = (\lambda + x_1) f_1^{0,-3/2} g_2$, we have

$$
\begin{aligned}
w &= \int x_1 f_1^{0,-3/2} g_2 \Delta \, d\Omega + \lambda \int f_1^{0,-3/2} g_2 \Delta \, d\Omega \\
&= -\frac{2\lambda}{n-1} w' + \int f_1^{1,-3/2} g_2 \Delta \, d\Omega \\
&= -\frac{2\lambda}{n-1} w' + \int f_1^{2,-3/2} g_2 e^{-x_1/2} \Delta \, dx_1 \, d\Omega_2 \\
&= -\frac{2\lambda}{n-1} \omega' - 2 \int f_1^{2,-3/2} g_2 \frac{d}{dx_1} \{ e^{-x_1/2} \} \Delta \, dx_1 \, d\Omega_2 \\
&= -\frac{2\lambda}{n-1} w' + 2 \int \frac{d}{dx_1} \{ f_1^{2,-3/2} \delta \} e^{-x_1/2} g_2 \Delta_2 \, dx_1 \, d\Omega_2 .
\end{aligned}
$$

The last line is the result of integration by parts. The differentiation gives three terms, so that

$$
\begin{aligned}
w = &-\frac{2\lambda}{n-1} w' + 4 \int f_1^{0,-3/2} g_2 \Delta \, d\Omega - 3 \int f_1^{1,-5/2} g_2 \Delta \, d\Omega \\
&+ 2(n-2) \int \frac{x_1}{x_1 - x_2} f_1^{0,-3/2} g_2 \Delta \, d\Omega \qquad (6) \\
= &-\frac{(2\lambda + 8) w'}{n-1} - 3 \int f_1^{1,-5/2} g_2 \Delta \, d\Omega + 2(n-2) \int \frac{x_1}{x_1 - x_2} f_1^{0,-3/2} g_2 \Delta \, d\Omega .
\end{aligned}
$$

Investigating each of the above two integrals, we find

$$
\int f_1^{1,-5/2} g_2 \Delta \, d\Omega = \int f_1^{0,-3/2} g_2 \Delta \, d\Omega - \lambda \int f_1^{0,-5/2} g_2 \Delta \, d\Omega, \qquad (7)
$$

and

$$
\begin{aligned}
\int \frac{x_1}{x_1 - x_2} f_1^{0,-3/2} g_2 \Delta \, d\Omega &= \int \frac{x_1 (x_2 + \lambda)}{x_1 - x_2} f_1^{0,-3/2} f_2^{0,-3/2} g_3 \Delta \, d\Omega \\
&= \lambda \int \frac{x_1}{x_1 - x_2} f_1^{0,-3/2} f_2^{0,-3/2} g_3 \Delta \, d\Omega,
\end{aligned}
$$

because $x_1 x_2 / (x_1 - x_2)$ is antisymmetric. We can use the identity $x_1/(x_1 - x_2) + x_2/(x_2 - x_1) = 1$ and symmetry to integrate this last expression. We obtain

$$
\begin{aligned}
\int \frac{x_1}{x_1 - x_2} f_1^{0,-3/2} g_2 \Delta\Omega &= \lambda \int \frac{x_1}{x_1 - x_2} f_1^{0,-3/2} f_2^{0,-3/2} g_3 \Delta\Omega \\
&= \frac{\lambda}{2} \int f_1^{0,-3/2} f_2^{0,-3/2} g_3 \Delta\Omega .
\end{aligned} \qquad (8)
$$

We substitute (7) and (8) into (6), replacing the integrals with the expressions for w' and w'', and finally rescale $z = \lambda/2$ to obtain (5). Equation (1) gives $w(0) = K_{n+1,n-1}^{-1}(n-1)!$ and clearly $w(\infty) = 0$. The constant term in the pdf is then

$$
\frac{K_n}{K_{n+1,n-1}} \frac{\Gamma(n/2+1)}{\Gamma(3/2)} = \frac{n}{2^{n-1/2}} \frac{\Gamma(n)}{\Gamma(n/2)},
$$

and the theorem is proved. This proof was inspired by [3]. See §§ 9 and 10 for further applications of the techniques used.

Though the pdf given in Theorem 3.1 is readily computed, the distribution of $n\lambda_{\min}$ is far simpler as $n \to \infty$.

COROLLARY 3.1. *If M_n has the distribution $W(n, n)$, then as $n \to \infty$, $n\lambda_{\min}$ converges in distribution to a random variable whose* pdf *is given by*

$$f(x) = \frac{1+\sqrt{x}}{2\sqrt{x}} e^{-(x/2+\sqrt{x})}.$$

Proof. From Theorem 3.1, the pdf of $n\lambda_{\min}$ is

$$f_{n\lambda_{\min}}(x) = \frac{n^{1/2}}{2^{n-1/2}} \frac{\Gamma(n)}{\Gamma(n/2)} x^{-1/2} e^{-x/2} U\left(\frac{n-1}{2}, -\frac{1}{2}, \frac{x}{2n}\right).$$

We recall that x_n converges to x in distribution if, for all α, $\lim_{n\to\infty} P(x_n < \alpha) = P(x < \alpha)$. We obtain pointwise convergence of the pdfs on $(0, \infty)$ with the aid of Stirling's formula and the following limiting expression:

$$\lim_{n\to\infty} 2\pi^{-1/2}\Gamma\left(\frac{n+2}{2}\right) U\left(\frac{n-1}{2}, -\frac{1}{2}, \frac{x}{2n}\right) = (1+\sqrt{x})e^{-\sqrt{x}},$$

which is a valid variation of equation 13.3.3 in [2].

In Fig. 3.1, we illustrate the speed of this convergence. We plot the ratio of the pdf of $n\lambda_{\min}$ for $n = 10$ against the function given by Corollary 3.1. We do the same for $n = 50$. Note for $n = 50$ the ratio is nearly 1 throughout the whole interval shown.

COROLLARY 3.2. *If M_n has the distribution $W(n, n)$, then as $n \to \infty$,*

$$E(\log(n\lambda_{\min})) \to -1.68788\cdots.$$

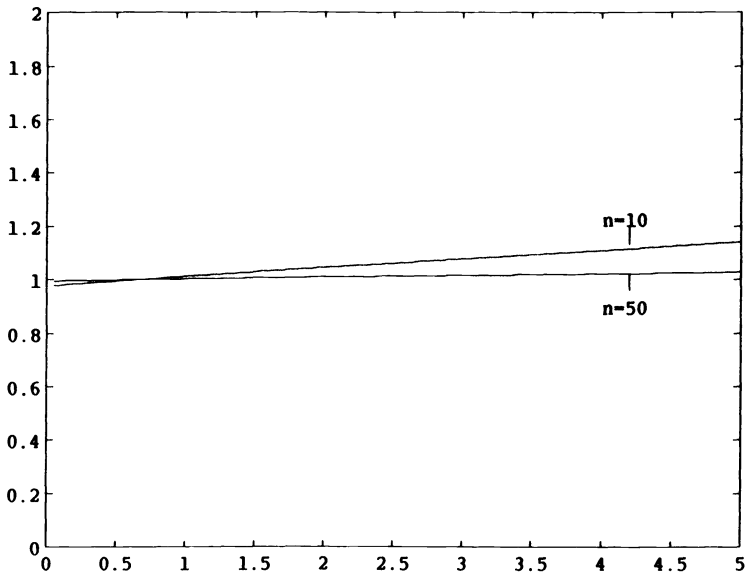


FIG. 3.1. *Speed of convergence of the* pdf *of $n\lambda_{\min}$.*

Proof. In light of the previous corollary and proper convergence of the integrals, the number we seek is

$$\int_0^\infty \log x \frac{1+\sqrt{x}}{2\sqrt{x}} e^{-(x/2+\sqrt{x})}\, dx.$$

This integral can be manipulated into

$$-2\gamma - 2e^{1/2}\int_1^\infty \frac{e^{-1/2y^2}}{y+1}\, dy$$

via a change of variables and equation 4.331.1 in [10], but we know of no simpler form. In this form, however, numerical integration is trivial. $\gamma \approx 0.5772$ in Euler's constant.

We now give the analogous results for complex matrices. The complex case turns out to be simpler.

THEOREM 3.2. *If M_n has the distribution $\tilde{W}(n, n)$, then the* pdf *of $\lambda_{\min}$ is given by*

$$f_{\lambda_{\min}}(\lambda) = \frac{n}{2} e^{-\lambda n/2}.$$

Proof. Let $f_{\lambda_{\min}}(\lambda)$ be the pdf of $\lambda_{\min}$. From (1) we have

$$f_{\lambda_{\min}}(\lambda) = \tilde{K}_{n,n} e^{-\lambda/2} \int_{R_\lambda} \exp\left(-\frac{1}{2}\sum_{i=1}^{n-1} \lambda_i\right) \prod_{i<j} (\lambda_i - \lambda_j)\, d\lambda_1 \cdots d\lambda_{n-1}.$$

By making the transformation $x_i = \lambda_i - \lambda$, we may conclude that $f_{\lambda_{\min}}(\lambda) = ce^{-n\lambda/2}$ for some constant c.

COROLLARY 3.3. *If M_n has the distribution $\tilde{W}(n, n)$, then for all n, $n\lambda_{\min}$ has the distribution χ_2^2.*

Although this corollary immediately follows from the theorem, we might only have guessed it immediately for $n = 1$. This result may be observed experimentally in Fig. 3.2, where we have computed $n\lambda_{\min}$ for 1000 matrices, each 100×100. After sorting these 1000 numbers, let η_i denote the ith value obtained. In Fig. 3.2, we plot η_i versus i/n. This gives the empirical fraction that is less than or equal to η_i. Note that this empirical cumulative density function (cdf) (also known as the empirical distribution function) wiggles around the theoretical cdf plotted as a solid line.

COROLLARY 3.4. *If M_n has the distribution $\tilde{W}(n, n)$, then for all n,*

$$E(\log n\lambda_{\min}) = \log 2 - \gamma = 0.11593\cdots.$$

Proof. We can use equation 4.352 in [10] to compute the appropriate integral.

4. The largest eigenvalue of $W(m, n)$ and $\tilde{W}(m, n)$. In this section we discuss the largest eigenvalue, $\lambda_{\max}$, of $W(n, n)$ and $\tilde{W}(n, n)$, but it requires little extra effort to consider a more general case. Specifically, consider a sequence of Wishart matrices $W(m_n, n)$ or $\tilde{W}(m_n, n)$ such that $m_n/n \to y$ as $n \to \infty$. Loosely speaking, we are looking at large matrices XX^T, where the ratio of number of rows to columns in X is roughly y. Clearly, $y = 1$ covers the cases of $W(n, n)$ and $\tilde{W}(n, n)$.

We start with a known result concerning the convergence in probability of the largest eigenvalues. As a reminder, to say $x_n \xrightarrow{p} x$ means for all $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|x - x_n| > \epsilon) = 0.$$

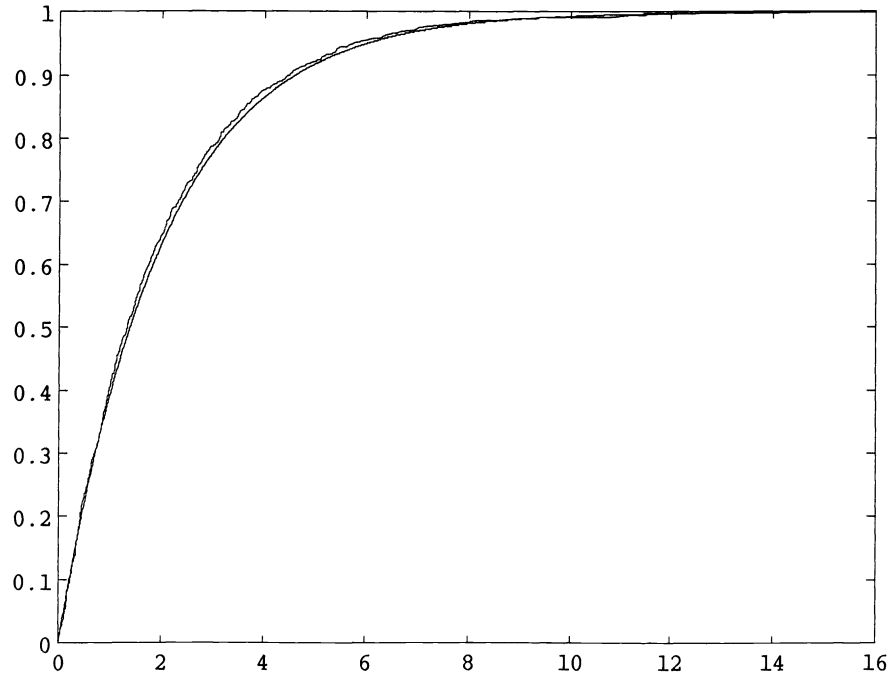


FIG. 3.2. *Theoretical and empirical* cdf *of* $n\lambda_{\min}$ *for* $W(n, n)$.

LEMMA 4.1. *If* M_n *has the distribution* $W(m_n, n)$, *where* $\lim_{n\to\infty} m_n/n = y$, $0 \leqq y < \infty$, *then*

$$(9) \qquad (1/n)\lambda_{\max} \xrightarrow{p} (1+\sqrt{y})^2 \quad \text{and for } 0 \leqq y \leqq 1, \quad (1/n)\lambda_{\min} \xrightarrow{p} (1-\sqrt{y})^2.$$

Proof. A stronger result (almost sure convergence) can be found in [17].

It is interesting to check Lemma 4.1 experimentally. When we take $y = 1$, the lemma states that, $(1/n)\lambda_{\max}$ converges in probability to 4. With $n = 100$, we computed $\lambda_{\max}/n$ for 1000 matrices. In Fig. 4.1, we plot the empirical cumulative density function (cdf), which is quite close to a step function with step at 4.

We would like to be able to readily conclude from Lemma 4.1 that

$$E(\log \lambda_{\max}/n) \to \log(1+\sqrt{y})^2.$$

It would be that simple if the logarithm were a bounded function; however, since $\log x$ has singularities at zero and infinity, we must carefully investigate the convergence at the singularities. To be precise, we must show that the sequence of random variables $\log \lambda_{\max}/n$ is uniformly integrable [5]. In the following lemma we estimate the pdf.

LEMMA 4.2. *If* M *has the distribution* $W(m, n)$, *then the* pdf $f_{\lambda_{\max}}(x)$ *satisfies*

$$(10) \qquad f_{\lambda_{\max}}(x) \leqq \frac{K_{n,m}}{K_{n-1,m-1}} x^{(n+m-3)/2} e^{-x/2} = \frac{\pi^{1/2} 2^{(1-n-m)/2}}{\Gamma(n/2)\Gamma(m/2)} x^{(n+m-3)/2} e^{-x/2}.$$

Proof. This was shown for $m = n$ in [22] by manipulating the expression (1). The same techniques work in the general case.

We can now prove the result that we expect.

PROPOSITION 4.1. *If* M_n *satisfies the hypotheses of Lemma* 4.1 *then* $E(\log \lambda_{\max}) = \log n + \log(1+\sqrt{y})^2 + o(1)$ *as* $n \to \infty$.

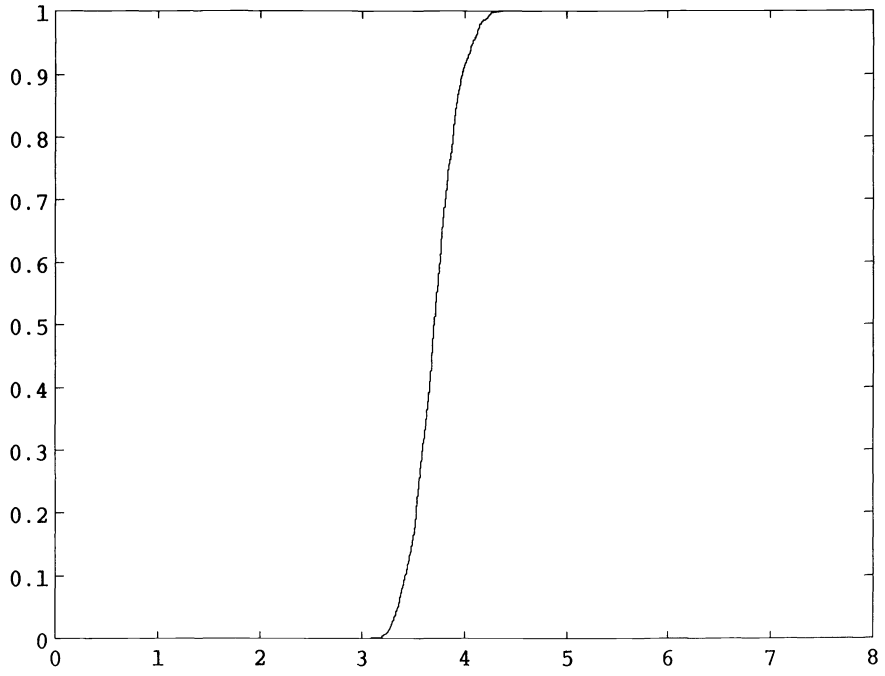


FIG. 4.1. *Empirical* cdf *of* $(1/n\lambda_{\max})$ *for* $W(n, n)$ $(n = 100)$.

Proof. Let σ denote $\lambda_{\max}/n$, and let $f_\sigma(x)$, $F_\sigma(x)$ be the corresponding probability density function and cumulative density function. We break up

$$E(\log \sigma) = \int_0^\infty \log x f_\sigma(x)\, dx$$

into three integrals:

$$\int_0^\varepsilon + \int_\varepsilon^r + \int_r^\infty$$

for values of ε and r depending on y, but not n. By Lemma 4.1, the middle integral approaches $\log (1 + \sqrt{y})^2$, and we proceed to show that the other integrals vanish in the limit.

Step 1. $\int_0^\varepsilon$.

We will need a fact that is also of independent interest. We have available another distribution of random matrices whose singular values are distributed exactly as that of $G(m, n)$. We perform a series of Householder transformations to obtain this distribution. (See [17] or [21] for details.) The conclusion is that if X has the distribution $G(m, n)$, then X is orthogonally similar to an $m \times n$ matrix

$$\begin{pmatrix} x_n & & & 0 \cdots 0 \\ y_{m-1} & x_{n-1} & & \vdots \\ & \ddots & \ddots & \vdots \\ & & y_1 \quad x_{n-(m-1)} & 0 \cdots 0 \end{pmatrix}, \tag{11}$$

where x_i^2 and y_i^2 are distributed as χ^2 variables with i degrees of freedom (i.e., χ_i^2). The elements here are all nonnegative and independent.

Let τ be the random variable defined by $(1/n)(x_n^2 + y_{m-1}^2)$. Considering the first column of (11), we have $\|M_n\| = \|X\|^2 = \lambda_{\max} \geqq x_n^2 + y_{m-1}^2$, i.e., $\sigma \geqq \tau$. It follows that $F_\sigma(x) \leqq F_\tau(x)$. Integrating by parts, we obtain

$$0 \geqq \int_0^1 \log x f_\sigma(x)\, dx = -\int_0^1 \frac{F_\sigma(x)}{x}\, dx \geqq -\int_0^1 \frac{F_\tau(x)}{x}\, dx = \int_0^1 \log x f_\tau(x)\, dx.$$

The terms $\log x\, F_\tau(x)$ and $\log x\, F_\sigma(x)$ produced by the integration by parts vanish as $x \to 0$. The former can be verified by using the fact that τ has the distribution $n^{-1}\chi^2_{n+m-1}$, and the latter follows from the former.

To complete the argument we take $m = m_n$, and let $k = n + m_n - 1$, so that τ has the distribution χ_k^2/n, and $f_\tau(x) = ((n/2)^{k/2}/\Gamma(k/2))x^{k/2-1}e^{-nx/2}$. Then,

$$0 \geqq \int_0^\varepsilon \log x f_\tau(x)\, dx \geqq \frac{(n/2)^{k/2}}{\Gamma(k/2)} \int_0^\varepsilon (\log x) x^{k/2-1} \approx \left(\frac{e\varepsilon}{1+y}\right)^{k/2}.$$

Here the $\approx$ indicates that only the exponential behavior is kept as $n \to \infty$. (Computing the asymptotics of this integral is routine but not obvious. A good reference is [4, Chap. 6].) By choosing any $\varepsilon < (1 + y)/e$, we have the desired result.

Step 2. $\int_r^\infty$.

For the singularity of the logarithm at ∞ we use Lemma 4.1, the fact that $f_\sigma(x) = nf_{\lambda_{\max}}(nx)$, and a standard asymptotic analysis.

For $r > 1 + y$,

$$\begin{aligned}\int_r^\infty f_\sigma(x) \log x\, dx &\leqq \int_r^\infty x f_\sigma(x)\, dx = \int_{rn/2}^\infty \left(\frac{4x}{n}\right) f_{\lambda_{\max}}(2x)\, dx \\ &\leqq \frac{(2/n)\pi^{1/2}}{\Gamma(n/2)\Gamma(m_n/2)} \int_{rn/2}^\infty x^{(n+m_n-1)/2} e^{-x}\, dx \\ &\approx (e^{-r}(er)^{1+y}y^{-y})^{n/2}.\end{aligned}$$

Here again, $\approx$ indicates that only the exponential behavior is kept as $n \to \infty$. By taking r (depending on y) sufficiently large, we conclude Step 2.

All of these results have analogues for the complex case.

LEMMA 4.3. *If M_n has the distribution $\tilde{W}(m_n, n)$, where $\lim_{n\to\infty} m_n/n = y$, $0 \leqq y < \infty$, then*

$$(1/n)\lambda_{\max} \xrightarrow{p} 2(1+\sqrt{y})^2 \text{ and for } 0 \leqq y \leqq 1, (1/n)\lambda_{\min} \xrightarrow{p} 2(1-\sqrt{y})^2. \tag{12}$$

PROPOSITION 4.2. *If M_n satisfies the hypotheses of Lemma* 4.3, *then $E(\log \lambda_{\max}) = \log n + \log 2(1 + \sqrt{y})^2 + o(1)$ as $n \to \infty$.*

The proofs are similar and are omitted, but we think it is of interest to mention the analogue of formula (11). If $\tilde{X}$ has the distribution $\tilde{G}(m, n)$, then $\tilde{X}$ is orthogonally similar to an $m \times n$ matrix

$$\begin{pmatrix} x_{2n} & & & & 0 \cdots 0 \\ y_{2(m-1)} & x_{2(n-1)} & & & \vdots \quad \vdots \\ & \ddots & \ddots & & \vdots \quad \vdots \\ & & y_2 & x_{2(n-(m-1))} & 0 \cdots 0 \end{pmatrix}, \tag{13}$$

where the notation is as in (11). From this we can immediately read that in the square complex case $\det \tilde{M}$ has the distribution $\chi^2_{2n}\chi^2_{2(n-1)}\cdots\chi^2_2$, while in the square real case it is well known (and can be seen from (11)) that $\det M$ has the distribution $\chi_n^2\chi_{n-1}^2\cdots\chi_1^2$.

5. The smallest eigenvalue of $W(m, n)$ and $\tilde{W}(m, n)$.

PROPOSITION 5.1. *If M_n satisfies the hypotheses of Lemma* 4.1 *and* $0 < y < 1$, *then* $E(\log \lambda_{\min}) = \log n + \log (1 - \sqrt{y})^2 + o(1)$.

Proof. As in the proof of Proposition 4.1, we must check that $\int_0^\varepsilon \log \lambda f_{\lambda_{\min}}(\lambda)\, d\lambda$ and $\int_r^\infty \log \lambda f_{\lambda_{\min}}(\lambda)\, d\lambda$ vanish as $n \to \infty$. We use the same notation as in the proof of Proposition 4.1 and abbreviate m_n as m:

$$
\begin{aligned}
&f_{\lambda_{\min}}(\lambda) \\
&= K_{n,m} \lambda^{(n-m-1)/2} e^{-\lambda/2} \int_{R_\lambda} \exp\left(-\sum_{i=1}^{m-1} \frac{\lambda_i}{2}\right) \prod_{i<j} (\lambda_i - \lambda_j) \prod_{i=1}^{m-1} (\lambda_i - \lambda) \lambda_i^{(n-m-1)/2}\, d\lambda_i \\
&\leqq K_{n,m} \lambda^{(n-m-1)/2} e^{-\lambda/2} \int_{R_0} \exp\left(-\sum_{i=1}^{m-1} \frac{\lambda_i}{2}\right) \prod_{i<j} (\lambda_i - \lambda_j) \prod_{i=1}^{m-1} \lambda_i^{(n-m+1)/2}\, d\lambda_i \\
&= \frac{K_{n,m}}{K_{n+1,m-1}} \lambda^{(n-m-1)/2} e^{-\lambda/2},
\end{aligned}
$$

and from (2),

$$
\frac{K_{n,m}}{K_{n+1,m-1}} = \pi^{1/2} 2^{-(n-m+1)/2} \Gamma\left(\frac{n+1}{2}\right) \Big/ \Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n-m+1}{2}\right) \Gamma\left(\frac{n-m+2}{2}\right).
$$

Let $\sigma = \lambda_{\min}/n$, so that $f_\sigma(x) = n f_{\lambda_{\min}}(nx)$. For $\varepsilon < 1 - y$,

$$
\begin{aligned}
0 \leqq \int_0^\varepsilon \log x f_\sigma(x)\, dx &\leqq \frac{K_{n,m}}{K_{n+1,m-1}} n^{(n-m+1)/2} \int_0^\varepsilon (\log x) x^{(n-m-1)/2} e^{-nx/2}\, dx \\
&\approx \left(\left(\frac{e\varepsilon}{n(1-y)^2}\right)^{1-y} e^{-\varepsilon}\right)^{n/2}.
\end{aligned}
$$

On the other hand, as in the proof of Proposition 4.1, $\sigma \leqq \tau$, which has the distribution χ^2_{n+m-1}. It then follows that $F_\sigma(x) \geqq F_\tau(x)$. For $r > 1$,

$$
\begin{aligned}
\int_r^\infty \log x f_\sigma(x)\, dx &= \log x (F_\sigma(x) - 1) \Big|_r^\infty + \int_r^\infty \frac{1 - F_\sigma(x)}{x} \\
&\leqq \log x (F_\sigma(x) - 1) \Big|_r^\infty - \log x (F_\tau(x) - 1) \Big|_r^\infty + \int_r^\infty \log x f_\tau(x)\, dx.
\end{aligned}
$$

The same kind of asymptotic analysis as above shows that as $n \to \infty$, each of the terms vanishes.

Of course, we have the complex result as well.

PROPOSITION 5.2. *If M_n satisfies the hypotheses of Lemma* 4.3 *and* $0 < y < 1$, *then* $E(\log \lambda_{\min}) = \log n + \log 2(1 - \sqrt{y})^2 + o(1)$.

6. Limiting condition number distributions and expected logarithms. We can now combine all the results of the previous section to describe the condition number distributions and the expected logarithms.

THEOREM 6.1. *If κ_n is the condition number of a matrix from the distribution $G(n, n)$, then κ_n/n converges in distribution to a random variable whose* pdf *is given by*

$$
f(x) = \frac{2x+4}{x^3} e^{-2/x - 2/x^2}.
$$

Moreover,

$$E(\log \kappa_n) = \log n + c + o(1) \approx \log n + 1.537$$

as $n \to \infty$.

Proof. From Lemma 4.1, we know $(1/n)\lambda_{\max} \xrightarrow{p} 4$ and Corollary 3.1 gives the limiting distribution for $n\lambda_{\min}$. The ratio of these quantities, κ_n^2/n^2, converges in distribution by a standard probability argument. The appropriate change of variables gives the limiting pdf of κ_n/n. The expected logarithm follows from Corollary 3.2 and Proposition 4.1.

THEOREM 6.2. *If* κ_n *is the condition number of a matrix from the distribution* $\tilde{G}(n, n)$, *then* κ_n/n *converges in distribution to a random variable whose* pdf *is given by*

$$f(x) = \frac{8}{x^3} e^{-4/x^2}.$$

Moreover,

$$E(\log \kappa_n) = \log n + \tfrac{1}{2}\gamma + \log 2 + o(1) \approx \log n + 0.982$$

as $n \to \infty$.

Proof. As in the proof of Theorem 6.1, the pdf follows from Lemma 4.3 and Corollary 3.3, and the expected logarithm follows from Corollary 3.4 and Proposition 4.2.

THEOREM 6.3. *If* κ_n *is the condition number of a matrix from the distribution* $G(m_n, n)$ *or* $\tilde{G}(m_n, n)$, *where* $\lim_{n\to\infty} m_n/n = y$ *and* $0 < y < 1$, *then* κ_n *converges in probability to* $(1 + \sqrt{y})/(1 - \sqrt{y})$. *Moreover,*

$$E(\log \kappa_n) = \log \frac{1+\sqrt{y}}{1-\sqrt{y}} + o(1)$$

as $n \to \infty$.

The convergence follows trivially from Lemma 4.1 and Lemma 4.2 and, of course, the statement could be strengthened to almost sure convergence. The expected logarithm follows from Propositions 4.1, 4.2, 5.1, and 5.2.

7. Exact expressions for $m = 2$. It is possible to integrate expressions (1) and (3) against the condition number to get the exact distributions of the condition numbers of real and complex $2 \times n$ matrices. We spare the reader the details and give only the results.

The pdf of the condition number of matrices that have the distribution $G(2, n)$ is given by

$$f_\kappa(x) = (n-1)2^{n-1} \frac{x^2-1}{(x^2+1)^n} x^{n-2}. \tag{14}$$

Similarly, when the matrices have the distribution $\tilde{G}(2, n)$, we have

$$f_{\tilde{\kappa}}(x) = 2\frac{\Gamma(2n)}{\Gamma(n)\Gamma(n-1)} \frac{x^{2n-3}(x^2-1)^2}{(x^2+1)^{2n}}. \tag{15}$$

We can use (14) and (15) to evaluate the integrals giving the expected condition numbers, and the result is the following theorem.

THEOREM 7.1. *If* X_n *has the distribution* $G(2, n)$, *then*

$$E(\log \kappa_{X_n}) = \frac{1}{2}\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right) \Big/ \Gamma\left(\frac{n}{2}\right).$$

If $\tilde{X}_n$ has the distribution $\tilde{G}(2, n)$, then

$$E(\log \kappa_{\tilde{X}_n}) = \log 2 + \frac{1}{2} - \sum_{k=2}^{n-1} \frac{1}{4^k}\binom{2k}{k}\frac{1}{k-1}.$$

We can also obtain the exact distribution for the smaller and the larger eigenvalues:

THEOREM 7.2. *If M_n has the distribution $W(2, n)$ and β denotes $(n-1)/2$, then*

$$f_{\lambda_{\min}}(\lambda) = K_{n,2}e^{-\lambda}(2\lambda^{\beta}e^{-\lambda/2} + 2^{\beta}(2\beta - \lambda)\Gamma(\beta, \lambda/2))$$

and

$$f_{\lambda_{\max}}(\lambda) = K_{n,2}e^{-\lambda/2}\lambda^{\beta-1}(2\lambda^{\beta}e^{-\lambda/2} - 2^{\beta}(2\beta - \lambda)\gamma(\beta, \lambda/2)).$$

A similar result for $\tilde{W}(2, n)$ could be calculated.

8. The tails of the condition number distributions. In the previous sections, we described the behavior of the condition numbers but said nothing about the probability that a matrix with a large condition number may appear. Here we will approximate the condition numbers for square matrices in order to get a sense of the tails of the distributions.

There are four condition numbers that we find interesting. Let κ and $\tilde{\kappa}$ denote the random variables, which are the 2-norm condition number of a matrix having the distribution $G(n, n)$ and $\tilde{G}(n, n)$, respectively. Since we are only considering $n \times n$ matrices, we omit the dependence on n in the notation. The other two condition numbers were introduced by Demmel [8]. Let $\|X\|_F$ denote the Frobenius norm of X, defined as $\sqrt{\sum_{i,j} X_{ij}^2} = \sqrt{\text{trace}\,(XX^T)}$. Demmel's condition number is defined by $\|X\|_F\|X^{-1}\|_2$. Let κ_D and $\tilde{\kappa}_D$ denote the random variables that are the Demmel condition number in the real and complex cases as above. We chart the condition numbers and relate them to the eigenvalues of the corresponding Wishart matrix in the table below.

$\kappa = \sqrt{\lambda_{\max}/\lambda_{\min}}$	$\tilde{\kappa} = \sqrt{\lambda_{\max}/\lambda_{\min}}$
$\kappa_D = \sqrt{\sum \lambda_i/\lambda_{\min}}$	$\tilde{\kappa}_D = \sqrt{\sum \lambda_i/\lambda_{\min}}$

In the tables that follow, we consistently use the above ordering: real versus complex in the columns, and 2-norm versus Demmel's norm in the rows.

The numbers in the table below are the values that the indicated expressions converge to in probability as $n \to \infty$.

	$W(n,n)$	$\tilde{W}(n,n)$
$\frac{1}{n}\lambda_{\max}$	4	8
$\frac{1}{n^2}\sum_{i=1}^{n}\lambda_i$	1	2

The first row is Lemmas 4.1 and 4.3. The second row is derived from the law of large numbers and the observation that the trace of a Wishart matrix has the $\chi^2_{n^2}$ distribution in the real case and the $\chi^2_{2n^2}$ in the complex case. Replacing these convergence results with equality, we define four approximate condition numbers:

$\kappa' = \sqrt{4n/\lambda_{\min}}$	$\tilde{\kappa}' = \sqrt{8n/\lambda_{\min}}$
$\kappa'_D = \sqrt{n^2/\lambda_{\min}}$	$\tilde{\kappa}'_D = \sqrt{2n^2/\lambda_{\min}}$

Directly from the definition of these condition numbers we have the following justification of our approximation.

LEMMA 8.1. *As* $n \to \infty$, κ/κ', κ_D/κ'_D, $\tilde{\kappa}/\tilde{\kappa}'$, *and* $\tilde{\kappa}_D/\tilde{\kappa}'_D$ *all converge in probability to* 1.

The approximate condition numbers only depend on $\lambda_{\min}$. Thus it becomes necessary to investigate the probability that $\lambda_{\min}$ is small.

LEMMA 8.2. *As* $\lambda \to 0$, $P(\lambda_{\min} < \lambda) \sim \sqrt{\lambda n}$ *if* M *has the distribution* $W(n, n)$ *and* $P(\lambda_{\min} < \lambda) \sim \lambda n/2$ *if* M *has the distribution* $\tilde{W}(n, n)$.

Proof. The real result comes from analyzing the formula given in Theorem 3.1. The complex result is trivial since $n\lambda_{\min}$ has the distribution χ_2^2 according to Corollary 3.3.

THEOREM 8.1. *As* $x \to \infty$,

$P(\kappa' > x) \sim 2n/x$	$P(\tilde{\kappa}' > x) \sim 4n^2/x^2$
$P(\kappa'_D > x) \sim n^{3/2}/x$	$P(\tilde{\kappa}'_D > x) \sim n^3/x^2$

Proof. Combine the small λ behavior described in Lemma 8.2 with the definitions of our condition numbers. The results follow from the obvious change of variables.

In one case we can compare our results with those known for the exact condition number. Demmel showed that for all n, $P(\tilde{\kappa}_D > x) \sim (n^3 - n)/x^2$ as $x \to \infty$, while we have $P(\tilde{\kappa}'_D > x) \sim n^3/x^2$ as $x \to \infty$. The difference is negligible for all but very small n.

9. All the eigenvalues of a Wishart matrix. We would like to describe the complete spectrum of a Wishart matrix. The m eigenvalues of a matrix from $W(m, n)$ and $\tilde{W}(m, n)$ are, of course, random, but what can we say about them? We have already mentioned their joint density function in (1) and (3), but this does not give much insight into the total picture. Here, we contrast three descriptions of the complete set of eigenvalues. The first two are well known and the third is, we believe, new.

(1) *Mode.* The m-tuple $(\lambda_1, \cdots, \lambda_m)$ that maximizes (1) or (3) (when there is a maximum) consists of the roots of the Laguerre polynomial

$$L_m^{(2(\alpha/\beta)-1)}(x/\beta),$$

where $\alpha = \frac{1}{2}(n - m - 1)$ and $\beta = 1$ in the real case, while $\alpha = n - m$ and $\beta = 2$ in the complex case.

(2) *Empirical distribution function.* Take a large Wishart matrix and plot the $(\lambda_i, i/n)$. The picture will be a curve the limiting form of which is well known and listed for reference in Propositions 9.1 and 9.2.

(3) *Expected characteristic polynomial.* The expected characteristic polynomials of Wishart matrices can be computed precisely. They are

$$(-\beta)^m m! L_m^{(n-m)}(t/\beta);$$

$\beta = 1$ in the real case and 2 in the complex case.

We now discuss these ideas in detail.

9.1. Mode. The mode is related to an electrostatic interpretation of the zeros of the classical polynomials given in [20]. Note that there is an infinite density in the real case when $m = n$ and $\lambda_m = 0$, so the formula does not apply.

9.2. Empirical distribution function. The empirical distribution function $W_M(x)$ of a matrix M is the fraction of eigenvalues of M that are less than or equal to x. One

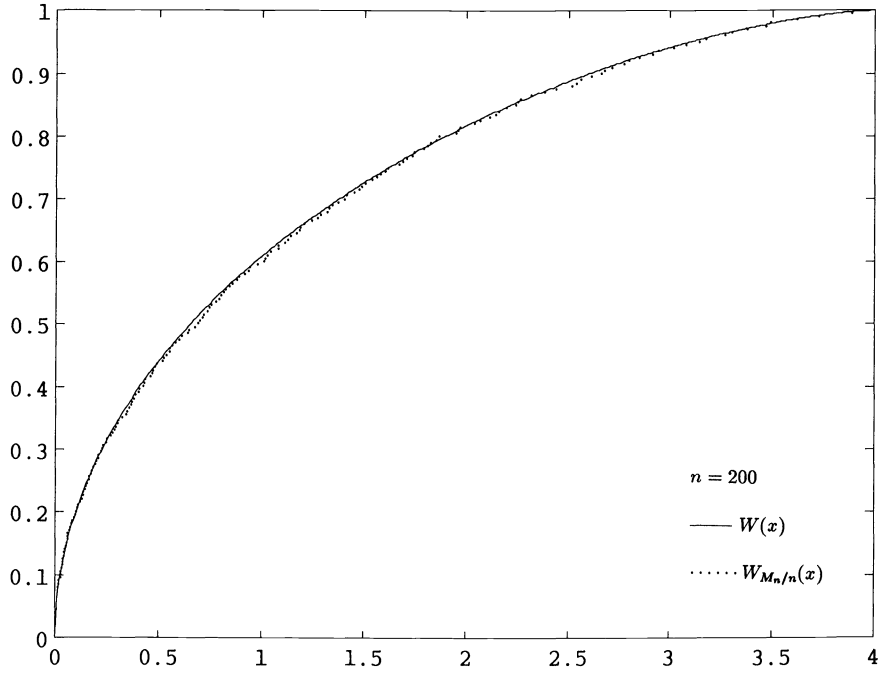


FIG. 9.1. *Empirical* cdf *of the eigenvalues of* $W(n, n)$.

way to view this is that if the eigenvalues are thought of as being chosen from a random sample, $W_M(x)$ is its empirical cdf. Computationally, we simply sort the eigenvalues and plot λ_i against i/n. We do this for a matrix M_n/n, where M_n was generated from the distribution $W(200, 200)$ and plot $W_{M_n}(x)$ in Figure 9.1 as a dotted line. It is well known that $W_{M_n}(x)$ converges almost surely to a limiting function as $n \to \infty$. $W(x)$ is plotted in Fig. 9.1 as a solid line.

PROPOSITION 9.1. *If* M_n *satisfies the conditions of Lemma* 4.1, *then* $W_{M_n/n}(x)$ *converges almost surely to a fixed function* $W(x)$ *as* $n \to \infty$. *If* $y = 1$, *this function satisfies* $W'(x) = (1/2\pi)((4 - x)/x)^{1/2}$ *for* $0 \leqq x \leqq 4$. *More generally, for* $0 < y \leqq 1$, *we have almost sure convergence to a fixed function satisfying*

$$W'(x) = \frac{\sqrt{(x - a(y))(b(y) - x)}}{2\pi yx}$$

for $a(y) < x < b(y)$, *where*

$$a(y) = (\sqrt{y} - 1)^2 \quad \text{and} \quad b(y) = (\sqrt{y} + 1)^2.$$

For $y > 1$ *the above result is modified by adding* $(1 - 1/y)\delta(x)$ *to* $W'(x)$.

Proof. This proposition and the one to follow was proved in [23] in a very general setting. Convergence in probability was proved earlier in [16]. Other more recent proofs can be found in [13] and [21]. These last two proofs are not as general but are quite elegant.

PROPOSITION 9.2. *If* $\tilde{M}_n$ *has the distribution* $\tilde{W}(m_n, n)$, *where* $\lim_{n\to\infty} m_n/n = y$ *and* $0 \leqq y < \infty$, *then* $W_{\tilde{M}/n}(x)$ *converges almost surely to a fixed function* $\tilde{W}(x)$ *as* $n \to \infty$. *If* $y = 1$, *this function satisfies* $\tilde{W}'(x) = (1/4\pi)((8 - x)/x)^{1/2}$ *for* $0 \leqq x \leqq 8$. *More generally,* $\tilde{W}(2x) = W(x)$, *as defined in Proposition* 9.1.

The source of the extra factor of 2 is simple. It is merely the variance of the elements of the matrices that are 1 in the real case but 2 in the complex case.

9.3. Characteristic polynomial. We can derive exactly the expected characteristic polynomial of a Wishart matrix. This could be thought of as the average of all the coefficients (which are of course symmetric functions of the eigenvalues) or as the average value of the characteristic polynomial at a given point. This is of interest here because the roots of the average polynomial deserve to be thought of as "typical" values for the eigenvalue.

Computing the expected characteristic polynomial is a special case of a multivariate integration of the form

$$\int_S f(\lambda_1, \cdots, \lambda_m)\Delta^k \, d\mu_1 \cdots d\mu_m,$$

where $d\mu_i = e^{-(1/2)\lambda_i}\lambda_i^\alpha d\lambda_i$, $\Delta = \prod_{i<j}(\lambda_i - \lambda_j)$, and the region of integration S is defined by $\lambda_1 \geqq \cdots \geqq \lambda_m \geqq 0$. Any expected value calculations involving the eigenvalues of Wishart matrices has exactly this form with $k = 1$ in the real case and $k = 2$ in the complex case. (See § 2.)

To compute the expected characteristic polynomial, take $f(\lambda_1, \cdots, \lambda_m) = \prod_{i=1}^p (t - \lambda_j)$, where t may be thought of as a variable. We make use of a recent result due to Aomoto [3].

LEMMA 9.1. *Let*

$$I_f = \int_{S_1} \prod_{i=1}^m (t-\lambda_i)\Delta^k \, d\nu_1 \cdots d\nu_m, \tag{16}$$

where $d\nu_i = \lambda_i^\alpha(1-\lambda_i)^\beta d\lambda_i$, *and the region of integration,* S_1 *is defined by* $1 \geqq \lambda_1 \geqq \cdots \geqq \lambda_m \geqq 0$. *Then*

$$\frac{I_f}{I_1} = \binom{\alpha' + \beta' + 2n}{n}^{-1} P_m^{(\alpha',\beta')}(1-2t), \tag{17}$$

where $P_m^{(\alpha',\beta')}$ *denotes a Jacobi polynomial,* $\alpha' = -1 + 2(\alpha+1)k$, $\beta' = -1 + 2(\beta+1)/k$ *and* $I_1 = \int_{S_1} \Delta^k \, d\nu_1 \cdots d\nu_m$.

This lemma is proved in [3]. The value of I_1 was first computed by Selberg in 1944, but his original paper is unavailable in many libraries. His results and argument, however, can be found in § 5.4 of [1]. We have derived an alternative proof to this lemma and to Lemma 9.2 by proving that the integrals satisfy the correct second-order differential equation for the Jacobi and Laguerre polynomials. This proof closely resembles the proof of Theorem 3.1.

LEMMA 9.2.

$$\int_S \prod_{i=1}^m (t-\lambda_i)\Delta^k \, d\mu_1 \cdots d\mu_m = c_{\alpha,m}^{(k)} L_m^{(\alpha')}\left(\frac{t}{k}\right), \tag{18}$$

where $L_m^{(\alpha')}$ *denotes a Laguerre polynomial,* $(c_{\alpha,m}^{(1)})^{-1} = (-1)^m\binom{m+\alpha}{m}K_{2\alpha+m+3,m}$ *and* $(c_{\alpha,m}^{(2)})^{-1} = (-1)^m\binom{m+\alpha}{m}\tilde{K}_{\alpha+m+1,m}$.

Proof. In (16) make the substitutions $\lambda_i \to \lambda_i/2\beta$ and $t \to t/2\beta$. The value of (17) becomes a multiple of

$$P_m^{(\alpha',\beta')}\left(1 - \frac{2t}{2\beta}\right) = P_m^{(\alpha',\beta')}\left(1 - \frac{2t}{k(\beta'+1)-2}\right).$$

To compute (18), let $\beta' \to \infty$. Using standard formulas about orthogonal polynomials, we can verify

$$\lim_{\beta' \to \infty} P_m^{(\alpha',\beta')}\left(1 - \frac{2t}{k(\beta'+1)-2}\right) = L_m^{(\alpha')}\left(\frac{t}{k}\right).$$

We get the constants $c_{\alpha,m}^{(k)}$ by setting $t = 0$ in (18). The right-hand side is $c_{\alpha,m}^{(k)}\binom{m+\alpha}{m}$. The left-hand integral is an integral of the expressions (1) and (3) up to a constant. Since (1) and (3) are joint density functions they integrate to 1. For a suitable choice of n we get the values (2) and (4). (Note that we computed the constant for $k = 1$ or 2 since we had (2) and (4) handy. We could have obtained $c_{\alpha,m}^{(k)}$ from scratch for all k, by evaluating the integral (18) when $t = 0$ by using a limiting process on the value of Selberg's integral.)

THEOREM 9.1. *Let $P_M(t) = \det(tI - M)$ be the characteristic polynomial of M. Then $E(P_M(t)) = (-1)^m m! L_m^{(n-m)}(t)$ if M has the distribution $W(m, n)$ and $E(P_M(t)) = (-2)^m m! L_m^{(n-m)}(t/2)$ if M has the distribution $\tilde{W}(m, n)$.*

Proof. Each of the expected values we are computing here has the form (18). In the real case $k = 1$ and $\alpha = (n - m - 1)/2$, so $\alpha' = n - m$. In the complex case, $k = 2$ and $\alpha = n - m$, so again $\alpha' = n - m$. The easy way to check that the constant is correct is to compare the highest coefficient of t, which is unity on both sides.

10. The probability density function of $\lambda_{\min}$ for $W(m, m + 3)$. The smallest eigenvalue of a matrix from $W(m, m + 1)$ behaves exactly like the one in $\tilde{W}(m, m)$, that is, $m\lambda_{\min}$ has the χ_2^2 distribution. The proof is similar to that of Theorem 3.2.

In fact, the pdf of $\lambda_{\min}$ for any matrix from $W(m, n)$ for $n - m$ odd or any matrix from $\tilde{W}(m, n)$ is given by

$$e^{-\lambda m/2} P(\lambda),$$

where P is a polynomial. This was pointed out in the real case in [15] and in fact can be seen directly from the integral.

To illustrate another application of Lemma 9.2, we derive the polynomial for the special case of $W(m, m + 3)$. A similar result is given in [15], where the distribution is expressed as a hypergeometric function of a matrix argument. The two results are in fact equivalent, but we give a more explicit expression.

THEOREM 10.1. *If M has the distribution $W(m, m + 3)$, then*

$$f_{\lambda_{\min}}(\lambda) = \frac{1}{2(m+1)} e^{-\lambda m/2} \lambda L_{m-1}^{(3)}(-\lambda).$$

Proof. From (1) we know that

$$f_{\lambda_{\min}}(\lambda) = K_{n,m} e^{-\lambda/2} \lambda \int_{S'} \prod_{i=1}^{m-1} (\lambda_i - \lambda) \Delta \prod_{i=1}^{m-1} \lambda_i \, d\lambda_1 \cdots d\lambda_{m-1},$$

where S' is defined by $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_{m-1} \geqq 0$. Letting $\lambda_i \to \lambda_i - \lambda$, we obtain

$$f_{\lambda_{\min}}(\lambda) = K_{n,m} e^{-\lambda m/2} \lambda \int_{S} \prod_{i=1}^{m-1} (\lambda_i + \lambda) \Delta \, d\mu_1 \cdots d\mu_{m-1}.$$

Here the notation is as in the previous section and $\alpha = 1$, so that $\alpha' = 3$. The conclusion follows from Lemma 9.2.

Acknowledgments. My sincere thanks go to Nick Trefethen, whose door was always open when I needed to describe an idea. His insightful comments have meant a great deal. I also wish to acknowledge the creators of an excellent software package, MATLAB, which made running experiments as easy as thinking of them.

REFERENCES

[1] G. E. ANDREWS, *q-Series: Their Development and Application in Analysis, Number Theory, Combinatorics, Physics, and Computer Algebra*, Regional Conference Series in Mathematics 66, American Mathematical Society, Providence, RI, 1986.

[2] M. ABRAMOWITZ AND I. A. STEGUN, EDS., *Handbook of Mathematical Functions*, Dover, New York, 1970.

[3] K. AOMOTO, *Jacobi polynomials associated with Selberg integrals*, SIAM J. Math. Anal., 18 (1987), 545–549.

[4] C. M. BENDER AND S. A. ORSZAG, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw–Hill, New York, 1978.

[5] P. BILLINGSLEY, *Probability and Measure*, John Wiley, New York, 1979.

[6] L. BLUM AND M. SHUB, *Evaluating rational functions: infinite precision is finite cost and tractable on average*, SIAM J. Comput., 15 (1986), pp. 384–398.

[7] J. E. COHEN, H. KESTEN, AND C. M. NEWMAN, EDS., *Random Matrices and Their Applications*, Contemporary Mathematics, Vol. 50, American Mathematical Society, Providence, RI, 1986.

[8] J. DEMMEL, *The probability that a numerical analysis problem is difficult*, Math. Comp., 50 (1988), pp. 449–480.

[9] S. GEMAN, *A limit theorem for the norm of random matrices*, Ann. Probab., 8 (1980), pp. 252–261.

[10] I. S. GRADSHTEYN AND I. W. RYZHIK, *Table of Integrals, Series, and Products*, Fourth edition, Academic Press, New York, 1965.

[11] R. D. GUPTA AND D. S. P. RICHARDS, *Hypergeometric functions of scalar matrix argument are expressible in terms of classical hypergeometric functions*, SIAM J. Math. Anal., 16 (1985), pp. 852–858.

[12] A. T. JAMES, *Distributions of matrix variates and latent roots derived from normal samples*, Ann. Math. Statist., 35 (1964), pp. 475–501.

[13] D. JONSSON, *Some limit theorems for the eigenvalues of a sample covariance matrix*, J. Multivariate Anal., 12 (1982), pp. 1–38.

[14] E. KOSTLAN, *Numerical linear algebra and multivariant analysis*, in preparation.

[15] P. R. KRISHNAIAH AND T. C. CHENG, *On the exact distribution of the smallest roots of the Wishart Matrix using zonal polynomials*, Ann. Inst. Statist. Math., 23 (1971), pp. 293–295.

[16] V. A. MARCENKO AND L. A. PASTUR, *Distributions of eigenvalues for some sets of random matrices*, Math. USSR-Sb., 1 (1967), pp. 457–483.

[17] J. W. SILVERSTEIN, *The smallest eigenvalue of a large-dimensional Wishart matrix*, Ann. Probab., 13 (1985), pp. 1364–1368.

[18] S. SMALE, *On the efficiency of algorithms of analysis*, Bull. Amer. Math Soc., 13 (1985), pp. 87–121.

[19] T. SUGIYAMA, *On the distribution of the largest latent root of the covariance matrix*, Ann. Math. Statist., 38 (1967), pp. 1148–1151.

[20] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1939.

[21] H. F. TROTTER, *Eigenvalue distributions of large hermitian matrices; Wigner's semi-circle law and a theorem of Kac, Murdock, and Szegö*, Adv. in Math., 54 (1984), pp. 67–82.

[22] J. VON NEUMANN AND H. H. GOLDSTINE, *Numerical inverting of matrices of high order*, in John von Neumann, Collected Works, Vol. 5: Design of Computers, Theory of Automata and Numerical Analysis, A. H. Taub, ed., Pergamon, New York, 1963.

[23] K. W. WACHTER, *The strong limits of random matrix spectra for sample matrices of independent elements*, Ann. Probab., 6 (1978), pp. 1–18.

[24] N. WEISS, G. W. WASILKOWSKI, H. WOŹNIAKOWSKI, AND M. SHUB, *Average condition number for solving linear equations*, Linear Algebra Appl., 83 (1986), pp. 79–102.

[25] S. WILKS, *Mathematical Statistics*, John Wiley, New York, 1967.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 4, October 1988

009

A NOTE ON THE NEWTON ITERATION FOR THE ALGEBRAIC EIGENVALUE PROBLEM*

MARIA CÉLIA SANTOS†

Abstract. This paper considers the Newton iteration for the algebraic eigenvalue problem $(\lambda I - A)x = 0$, $\Phi(x) = 1$, where Φ is a convex norming function that is not necessarily differentiable. The role usually played by the Fréchet or Gateaux derivatives will be performed by a choice of subgradients of Φ. Under very mild conditions on Φ, the local and Q-superlinear convergence of this extended Newton iteration are proved. The stability of the process is also investigated.

Key words. Newton's iteration, algebraic eigenvalue problem, subgradients

AMS(MOS) subject classification. 65F10

1. Introduction. Let A be an $n \times n$ real matrix. The Newton's iteration is a well-known and thoroughly studied method for the iterative determination of a real eigenvalue and a corresponding eigenvector of A (cf., e.g., [1]–[4]).

The real eigenvalue problem consists of finding a real number μ and a vector v of $\mathbb{R}^n$ satisfying

$$(\mu I - A)v = 0, \qquad \Phi(v) = 1, \tag{1.1}$$

where Φ: $\mathbb{R}^n \to \mathbb{R}$ is a norming function in $\mathbb{R}^n$ (usually Φ is a norm, an affine function, a quadratic form, etc.). If we define $F(\lambda, x)$ by

$$F(\lambda, x) := \begin{bmatrix} (\lambda I - A)x \\ \Phi(x) - 1 \end{bmatrix}, \tag{1.2}$$

for any $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$, then (1.1) is equivalent to $F(\mu, v) = 0$.

Now assume that Φ is differentiable and let $J(\lambda, x)$ denote the Jacobi matrix of $F(\lambda, x)$, namely

$$J(\lambda, x) := \begin{bmatrix} \lambda I - A & x \\ g_x & 0 \end{bmatrix}, \tag{1.3}$$

where we represent by g_x the gradient of Φ at x (gradients will be considered as row-vectors throughout). With this notation, the Newton iteration has the following formal presentation:

$$\begin{bmatrix} x_{k+1} \\ \lambda_{k+1} \end{bmatrix} := \begin{bmatrix} x_k \\ \lambda_k \end{bmatrix} - J(\lambda_k, x_k)^{-1} F(\lambda_k, x_k). \tag{1.4}$$

This method is started with any pair (λ_0, x_0) and then the iteration proceeds according to (1.4), where it is assumed that $J(\lambda_k, x_k)$ is invertible.

The Newton iteration (1.4) has been extensively studied in the literature, under the assumption that Φ is G-differentiable at each x_k [1]–[3], [5], [8]. However, even in theoretical discussions, the algorithm has not been handled, assuming no differentiability at some iterate x_k. As a simple example, let us take Φ as $\|\cdot\|_\infty$, which is frequently used. Moreover, let us suppose that at some step k the computed x_k has (up to machine accuracy) more than one component of maximum modulus. This means that Φ is not differ-

* Received by the editors May 25, 1987; accepted for publication (in revised form) March 22, 1988.
† Departamento de Matemática, Universidade de Coimbra, Coimbra, Portugal.

entiable at x_k. In this situation, procedure (1.4) must stop. However, the subdifferentiability of $\|\cdot\|_\infty$ at each $x \in \mathbb{R}^n$ provides a natural way to overcome such a situation. Namely, the role of g_x may be played by a subgradient of $\|\cdot\|_\infty$ at x. To be more specific, if $x_k = (x_k^{(1)}, \cdots, x_k^{(n)})^T$ is nonzero, a subgradient of $\|\cdot\|_\infty$ at x_k is any vector $g = (g^{(1)}, \cdots, g^{(n)})$ such that $|g^{(1)}| + \cdots + |g^{(n)}| = 1$ and $\operatorname{sgn}(g^{(i)}) = \operatorname{sgn}(x_k^{(i)})$ for $i = 1, \cdots, n$ (for a real ξ we let $\operatorname{sgn}(\xi) = 1$ if $\xi > 0$, $\operatorname{sgn}(\xi) = -1$ if $\xi < 0$, and $\operatorname{sgn}(\xi) = 0$ if $\xi = 0$); if $x_k = 0$, g is a subgradient of $\|\cdot\|_\infty$ at x_k if and only if $|g^{(1)}| + \cdots + |g^{(n)}| \leqq 1$. Thus we select any such g to play the role of g_{x_k} and the computation of x_{k+1} is carried out according to (1.3) and (1.4).

The nondifferentiable norm $\Phi(x) := \|x\|_1$ is also often used. Here g is a subgradient of $\|\cdot\|_1$ at $x \neq 0$ if and only if $\|g\|_\infty = 1$ and $g^{(i)} = \operatorname{sgn}(x^{(i)})$ if $x^{(i)} \neq 0$. Moreover, g is a subgradient of $\|\cdot\|_1$ at $x = 0$ if and only if $\|g\|_\infty \leqq 1$. Of course, if it happens that x_k has (up to machine accuracy) at least one zero component, then $\|\cdot\|_1$ is not differentiable at x_k, i.e., we have more than one subgradient at x_k. Then we may let any such subgradient play the role of g_{x_k} in (1.3) and (1.4).

Generally speaking, the purpose of this note is to show that, as far as the local behaviour is concerned, the convergence properties of the extended Newton method that we have just roughly described are the same as those of the usual Newton method for differentiable functions. More precisely, under the basic assumption that Φ is convex, we shall prove a point of attraction theorem and the Q-superlinear convergence of the extended method (§ 3). The local stability is established in § 4. We retrieve, as we should, the classical point of attraction and stability theorems when Φ is differentiable. Since we extend the class of cases to which iteration (1.4) is applicable, it is obvious that we cannot expect a lesser degree of complexity than that of the usual Newton iteration with differentiable norming functions.

2. Preliminary considerations. In the sequel, $\Phi\colon \mathbb{R}^n \to \mathbb{R}$ denotes a real convex function. A vector g in $\mathbb{R}^n$ is called a *subgradient of* Φ *at* $v \in \mathbb{R}^n$, if the following holds:

$$\Phi(x) - \Phi(v) \geqq \langle g, x - v \rangle$$

for any x in $\mathbb{R}^n$. Here and throughout we represent the usual inner product in $\mathbb{R}^n$ by $\langle x, y \rangle$. The *subdifferential of* Φ *at* v is the set of all subgradients of Φ at v; it is denoted by $\partial\Phi(v)$.

For this concept and relevant results we refer the reader to [6].

Recall that $\partial\Phi(v)$ is a nonempty, compact, convex set; it reduces to a singleton if and only if Φ is differentiable at v.

For future reference we set forth the following comment.

Remark 2.1. Denote by $\Phi'(x; y)$ the one-sided directional derivative of Φ at x, with respect to y. By [6, Thm. 23.4] it is easy to see that $\{\langle g, y \rangle : g \in \partial\Phi(x)\}$ is the real closed interval whose extremes are $-\Phi'(x; -y)$ and $\Phi'(x; y)$.

The *subdifferential of the function* F given in (1.2) *at* (λ, x) will be denoted by $\partial F(\lambda, x)$ and is defined as the set of matrices

$$J(\lambda, x, g) := \begin{bmatrix} \lambda I - A & x \\ g & 0 \end{bmatrix}, \tag{2.1}$$

where g runs over the set $\partial\Phi(x)$. The matrix (2.1) will be referred to as a *Jacobi matrix of* F *at* (λ, x).

DEFINITION 2.2. We say that a convex function $\Phi\colon \mathbb{R}^n \to \mathbb{R}$ is a *norming function* if the following conditions hold:

(i) Every nonzero $x \in \mathbb{R}^n$ is *positively Φ-normalizable*, that is, there exists $\alpha > 0$ such that $\Phi(\alpha x) = 1$.

(ii) $\langle g, x \rangle \neq 0$, for any nonzero x and any $g \in \partial\Phi(x)$.

PROPOSITION 2.3. *For a convex function $\Phi: \mathbb{R}^n \to \mathbb{R}$ the following are equivalent*:

(a) *Φ is a norming function*;

(b) *Φ has a strict minimum at $x = 0$ and $\Phi(0) < 1$.*

Proof. For any norming function Φ condition (i) implies that the level set $B_\Phi := \{x: \Phi(x) \leqq 1\}$ is a compact nonempty set. Therefore Φ attains its minimum at a certain point of B_Φ. It is well known that $\Phi(x_0)$ is the minimum of Φ if and only if $0 \in \partial\Phi(x_0)$. Hence, condition (ii) implies that x_0 is the unique minimizing point of Φ. Therefore (a) implies (b).

Conversely, assume that (b) holds and let $x \neq 0$. For any $g \in \partial\Phi(x)$ we have $0 < \Phi(x) - \Phi(0) \leqq \langle g, x \rangle$, and therefore (ii) is true. On the other hand, (i) follows easily from the three following facts: (1) the continuity of Φ; (2) $\Phi(0) < 1$; (3) $\lim \Phi(tx) = +\infty$ as $t \to +\infty$. □

Let us briefly discuss the invertibility of matrices of type (2.1). This problem has been considered for example in [8], so we content ourselves with the following statement without proof.

PROPOSITION 2.4. *If μ is a real eigenvalue of A and v is a corresponding eigenvector, then $J(\mu, v, g)$ is nonsingular if and only if μ has multiplicity one and $\langle g, v \rangle \neq 0$.* □

Note that Proposition 2.4 does not require that $g \in \partial\Phi(v)$.

We say that the subdifferential of F at (λ, x) is *unconditionally invertible* if the Jacobi matrices (2.1) are invertible, for all $g \in \partial\Phi(x)$.

As an easy consequence of the above proposition, we have the following corollary.

COROLLARY 2.5. *Let Φ be a norming function and let (μ, v) be a real eigenpair of A. Then the subdifferential $\partial F(\mu, v)$ is unconditionally invertible if and only if μ is simple.* □

PROPOSITION 2.6. *Let $\Phi: \mathbb{R}^n \to \mathbb{R}$ be a convex function. The set of the points (λ, x) for which $\partial F(\lambda, x)$ is unconditionally invertible is an open set of $\mathbb{R}^{n+1}$.*

Proof. Let $\partial F(\lambda, x)$ be unconditionally invertible. Seeking a contradiction, we assume that there exists a sequence (λ_k, x_k), converging to (λ, x) such that $\partial F(\lambda_k, x_k)$ is *not* unconditionally invertible, for $k = 1, 2, \cdots$. This means that, for any k, there exists $g_k \in \partial\Phi(x_k)$ such that the matrix $J(\lambda_k, x_k, g_k)$ is singular. Since the set

$$S := \{x, x_1, \cdots, x_k, \cdots\}$$

is compact, then $\partial\Phi(S) := \cup \{\partial\Phi(y): y \in S\}$ is compact as well (cf. [6, Thm. 24.7]). Therefore, there exists a subsequence of (g_k) converging to an element $\bar{g}$ of $\partial\Phi(x)$ (cf. [6, Thm. 24.4]). Thus $J(\lambda, x, \bar{g})$ is singular. This contradicts the unconditional invertibility of $\partial F(\lambda, x)$. □

Now, assume again we are given an $n \times n$ real matrix A and a norming function Φ. The Newton iteration for the eigenvalue problem will be carried out according to the following scheme.

ALGORITHM 2.7. The iteration starts with a real number λ_0 and a vector $x_0 \in \mathbb{R}^n$. For $k = 0, 1, \cdots$ we proceed inductively as follows:

(N.1) Choose g_k in the set $\partial\Phi(x_k)$;

(N.2) If $J(\lambda_k, x_k, g_k)$ is singular, then the algorithm halts;

(N.3) If $J(\lambda_k, x_k, g_k)$ is nonsingular, then define (λ_{k+1}, x_{k+1}) by

$$\begin{bmatrix} x_{k+1} \\ \lambda_{k+1} \end{bmatrix} := \begin{bmatrix} x_k \\ \lambda_k \end{bmatrix} - J(\lambda_k, x_k, g_k)^{-1} F(\lambda_k, x_k).$$

The equation of (N.3) is obviously equivalent to the system

$$(\lambda_k I - A)x_{k+1} = (\lambda_k - \lambda_{k+1})x_k,$$
$$\langle g_k, x_{k+1}\rangle = \langle g_k, x_k\rangle + 1 - \Phi(x_k). \tag{2.2}$$

Remark 2.8. Assume that in a certain subset W of $\mathbb{R}^n$, we have

$$\langle g_x, y\rangle = \Phi(y) \tag{2.3}$$

for all $x, y \in W$ and $g_x \in \partial\Phi(x)$. It is clear that this condition is equivalent to the fact that Φ is *locally linear* in W, that is, for any $x = (x_1, \cdots, x_n)^T \in W$,

$$\Phi(x) = \alpha_1 x_1 + \cdots + \alpha_n x_n,$$

where $\alpha_1, \cdots, \alpha_n$ are constants (the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are examples of such functions).

If (2.3) holds and the iterates belong to W, then Algorithm 2.7 is equivalent to the well-known Wielandt iteration (cf. [3]).

3. Convergence theorems. We point out that in Algorithm 2.7 the pair (λ_{k+1}, x_{k+1}) depends on (λ_k, x_k) and on the choice of the subgradient made in step (N.1).

Let us denote by $\mathscr{C}(\lambda_0, x_0)$ the set of all sequences that may be generated by Algorithm 2.7 and having the same starting pair (λ_0, x_0). If Φ is differentiable, then the set $\mathscr{C}(\lambda_0, x_0)$ obviously consists of exactly one sequence. It is worth noting that some of the sequences of $\mathscr{C}(\lambda_0, x_0)$ may be *finite*, because the algorithm may very well stop at step (N.2). According to convention, the finite sequences of $\mathscr{C}(\lambda_0, x_0)$ are *not convergent.*

DEFINITION 3.1. We say that (λ_0, x_0) is an *unconditional pair with respect to* (μ, v) if any sequence of $\mathscr{C}(x_0, \lambda_0)$ converges to (μ, v).

From now on, we will use Φ to represent a norming function and A an $n \times n$ real matrix.

ATTRACTION THEOREM 3.2. *Let μ be a real, simple eigenvalue of A, and let v be a corresponding eigenvector such that $\Phi(v) = 1$. Then there exists a neighborhood of (μ, v) whose elements are unconditional pairs with respect to (μ, v). The process has superlinear convergence.*

The proof of this theorem follows a traditional pattern (cf., e.g., [2]), except for a few technical details contained in the following lemmas.

The first lemma is a sort of mean value theorem for convex functions.

LEMMA 3.3. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a convex function. For any vectors $x, u \in \mathbb{R}^n$, $x \neq u$, there exist $\hat{x} \in \mathbb{R}^n$ and $\hat{g} \in \partial f(\hat{x})$ such that we have the following*:

(i) *$\hat{x}$ is of the form $\hat{x} = x + \tau(x - u)$, with $0 < \tau < 1$;*

(ii) $f(x) - f(u) = \langle \hat{g}, x - u\rangle$.

Proof. The lemma is easy to prove in the case $n = 1$. If $x = u$ the lemma is trivial. Next, we prove the lemma in the general case, assuming that $x \neq u$. Let us define $y := \|x - u\|^{-1}(x - u)$ and consider the convex function $\psi(\theta) := f(u + \theta y)$. Since the lemma is true for $n = 1$, there exist $\alpha \in \mathbb{R}$ and $\gamma \in \partial\psi(\alpha)$, such that $0 < \alpha < \|x - u\|$ and

$$\psi(\|x-u\|) - \psi(0) = \gamma\|x-u\|. \tag{3.1}$$

On the other hand, if we define $\tau := \alpha\|x - u\|^{-1}$ and $\hat{x} := u + \tau(x - u)$, then we can easily see that

$$\partial\psi(\alpha) = \{\langle g, y\rangle : g \in \partial(\hat{x})\}.$$

Therefore there exists $\hat{g} \in \partial f(\hat{x})$ such that $\gamma = \langle \hat{g}, y \rangle$. With this notation, (3.1) implies (ii). □

LEMMA 3.4. *Let $f: \mathbb{R}^n \to \mathbb{R}$ be a convex function, and let u be any element of $\mathbb{R}^n$. Then, for any $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$|f(x) - f(u) - \langle g, (x-u) \rangle| \leqq \varepsilon \|x-u\|,$$

for all x such that $\|x-u\| < \delta$ and all $g \in \partial f(x)$.

Proof. We will get a contradiction from the assumption that there exist $\varepsilon_1 > 0$ and sequences of vectors (x_m) and (g_m) satisfying the following:

$$\begin{aligned} &(x_m) \text{ converges to } u, g_m \in \partial f(x_m), \\ &|f(x_m) - f(u) - \langle g_m, x_m - u \rangle| > \varepsilon_1 \|x_m - u\| \quad \text{for all } m. \end{aligned} \tag{3.2}$$

Denote by y_m the vector $\|x_m - u\|^{-1}(x_m - u)$. We may assume, without loss of generality, that (y_m) is a convergent sequence. Denote by y the limit of (y_m).

By the previous lemma, for each m there exists $\hat{x}_m$ of the form

$$\hat{x}_m = u + \tau_m(x_m - u) \quad (0 < \tau_m < 1),$$

such that

$$f(x_m) - f(u) = \langle \hat{g}_m, y_m \rangle \|x_m - u\| \tag{3.3}$$

for some $\hat{g}_m \in \partial f(\hat{x}_m)$.

As $(\hat{x}_m)$ converges to u, by [6, Thm. 24.6], for any $\varepsilon > 0$, there exists m_0 such that

$$\partial f(x_m) \cup \partial f(\hat{x}_m) \subset \partial f(u)_y + \varepsilon B, \tag{3.4}$$

for all $m \geqq m_0$. In (3.4), B is the Euclidean unit ball and $\partial f(u)_y$ is the set of vectors $h \in \partial f(u)$ such that

$$\langle h, y \rangle = \sup \{ \langle g, y \rangle : g \in \partial f(u) \}. \tag{3.5}$$

Thus there exist vectors h_m and $\hat{h}_m \in \partial f(u)_y$, depending on ε, satisfying

$$\|g_m - h_m\| < \varepsilon \quad \text{and} \quad \|\hat{g}_m - \hat{h}_m\| < \varepsilon,$$

for all $m \geqq m_0$. By (3.5), $\langle h_m, y \rangle = \langle \hat{h}_m, y \rangle$. Therefore, taking (3.3) into account, we easily obtain the following inequality:

$$\begin{aligned} &|f(x_m) - f(u) - \langle g_m, x_m - u \rangle| \\ &\quad \leqq \|x_m - u\| [\, |\langle \hat{g}_m, y_m - y \rangle| + |\langle \hat{g}_m - \hat{h}_m, y \rangle| \\ &\qquad + |\langle h_m - g_m, y \rangle| + |\langle g_m, y - y_m \rangle| \,]. \end{aligned} \tag{3.6}$$

By a compactness property of subdifferentials (cf. [6, Thm. 24.7]), the sequences (g_m) and $(\hat{g}_m)$ are uniformly bounded, because (x_m) and $(\hat{x}_m)$ both converge to u. Moreover, $(g_m - h_m)$, $(\hat{g}_m - \hat{h}_m)$, and $(y_m - y)$ converge to zero. Therefore, the bracketed term displayed in (3.6) tends to zero. This contradicts (3.2). □

Proof of Theorem 3.2. Let us define $z := (\lambda, x)$ and $z^* := (\mu, v)$. Denote by $T(z, g)$ the extended Newton operator, given by

$$T(z, g) = z - J(\lambda, x, g)^{-1} F(\lambda, x).$$

Observe that, by Propositions 2.4 and 2.6, there exists a compact neighborhood V of z^* such that $J(\lambda, x, g)$ is nonsingular for any (λ, x) in V and any g in $\partial \Phi(x)$. The compact-

ness of V [6, Thm. 24.7] and a standard continuity argument yield a constant $\beta > 0$ such that

$$\|J(\lambda, x, g)^{-1}(z - z^*)\| \leqq \beta \|z - z^*\|$$

for any $z \in V$. With a few calculations it is easy to prove that

$$\|T(z,g) - z^*\| \leqq \beta \|z - z^*\|^2 + \beta |\Phi(x) - \Phi(v) - \langle g, x - v \rangle| \tag{3.7}$$

for all z in V and g in $\partial\Phi(x)$. Therefore, using Lemma 3.3, we can show that for any $\varepsilon > 0$ there exists a neighborhood V_ε of z^* such that

$$\|T(z,g) - z^*\| \leqq \varepsilon \|z - z^*\|.$$

The proof can now be completed in the same manner as in [2]. □

In the following comments we denote a simple eigenvalue of A by μ, and v and w represent the corresponding eigenvectors, such that $\Phi(v) = \Phi(w) = 1$ and $w = -\sigma v$ with $\sigma > 0$.

Let $\psi\colon \mathbb{R} \to \mathbb{R}$ be a convex norming function and let α (respectively, β) be the unique positive (respectively, negative) real number such that $\psi(\alpha) = 1$ (respectively, $\psi(\beta) = 1$). We may consider the Newton iteration given by

$$t_{p+1} := t_p - \frac{\psi(t_p) - 1}{\gamma_p}, \qquad p = 0, 1, 2, \cdots. \tag{3.8}$$

This iteration is initialized with a real number $t_0 \neq 0$ and, at the pth step, we choose γ_p in $\partial\psi(t_p)$ (this means that $\psi'_-(t_p) \leqq \gamma_p \leqq \psi'_+(t_p)$, where ψ'_- and ψ'_+ are the left and right derivatives of ψ).

It is easily seen that

$$\text{If } t_0 > 0 \text{ then } \lim_p t_p = \alpha, \tag{3.9}$$

$$\text{If } t_0 < 0 \text{ then } \lim_p t_p = \beta. \tag{3.10}$$

As a matter of fact, if $t_0 > 0$ it follows by induction on $p \geqq 0$ that $\gamma_p \geqq 0$ and $t_{p+1} \geqq t_{p+2} \geqq \alpha$. Therefore (3.9) holds. We prove (3.10) in the same manner.

Now assume that we start Algorithm 2.7 with $x_0 := t_0 v$ and $\lambda_0 := \mu$, where $t_0 \neq 0$. It is easily seen that $\partial F(\mu, tv)$ is unconditionally invertible if $t \neq 0$. Solving (2.2) by induction, we find that any sequence (λ_p, x_p) of the set $\mathscr{C}(\lambda_0, x_0)$ has the form

$$\lambda_p = \mu \quad \text{and} \quad x_p = t_p v \quad \text{for } p = 0, 1, 2, \cdots,$$

where t_p is given by (3.8) for ψ defined as $\psi(t) := \Phi(tv)$. Therefore we have:

(3.11) For $t_0 \neq 0$, the pair $(\mu, t_0 v)$ is unconditional with respect to either (μ, v) (if $t_0 > 0$) or (μ, w) (if $t_0 < 0$).

Next we consider the case where Algorithm 2.7 is started with (λ_0, x_0), $\lambda_0 = \mu$. It is clear that (cf. [8]) the nonsingularity of $J(\mu, x_0, g_0)$ is equivalent to the following assumption:

(3.12) $\langle g_0, v \rangle \neq 0$ and x_0 do not belong to the column space of $\mu I - A$.

If this assumption holds, formulae (2.2) for $k = 0$ give us a first iterate of the form $(\lambda_1, x_1) = (\mu, tv)$, where $t \cdot \langle g_0, v \rangle > 0$. Therefore by (3.11):

(3.13) If assumption (3.12) holds for any $g_0 \in \partial\Phi(x_0)$, then (μ, x_0) is unconditional with respect to (μ, v) or (μ, w) according to $\langle g_0, v \rangle > 0$ or $\langle g_0, v \rangle < 0$, respectively.

Note that in the above statement, the product $\langle g_0, v\rangle$ has a well-determined signal, because $\partial\Phi(x_0)$ is a connected compact set.

Our final claim deals with the case when we start Algorithm 2.7 with an eigenvector of A.

(3.14) For all reals $t_0 \neq 0$ and $\lambda_0 \neq 0$, the infinite sequences of $\mathscr{C}(\lambda_0, t_0 v)$ converge to (μ, v) or to (μ, w), according to whether t_0 is positive or negative.

Proof. Assume $t_0 > 0$ (the case when $t_0 < 0$ is similar). If (λ_p, x_p) is an infinite sequence of $\mathscr{C}(\lambda_0, t_0 v)$, then $J(\lambda_p, x_p, g_p)$ is nonsingular for all p, where g_p is the subgradient chosen in step (N.1) of Algorithm 2.7. As x_0 is an eigenvector of A, the nonsingularity of $J(\lambda_0, x_0, g_0)$ implies $\langle g_0, v\rangle \neq 0$. Therefore, if we solve (2.2) for $k = 0$, we get

$$x_1 = t_1 v, \qquad \lambda_1 = \lambda_0 - \frac{(\lambda_0 - \mu)t_1}{t_0},$$

where t_1 is given by

$$t_1 := t_0 - \frac{\Phi(t_0 v) - 1}{\langle g_0, v\rangle}.$$

As we saw in the proof of (3.9), t_1 is positive. Therefore, by induction on p, we have

$$x_p = t_p v, \qquad \lambda_{p+1} = \lambda_p - (\lambda_p - \mu)t_{p+1}/t_p,$$

where t_{p+1} is given by (3.8), with $\psi(t) := \Phi(tv)$. By (3.8) we may conclude that $((\lambda_p, x_p))$ converges to (μ, v). □

Let $\Phi\colon \mathbb{R}^n \to \mathbb{R}$ be a convex function and $q > 0$. We say that $\partial\Phi$ is *weakly q-continuous at* v, if there exists a neighborhood W of v and a constant $\gamma \geqq 0$ such that:

(3.15) For any x in W and any g in $\partial\Phi(x)$, there exists h in $\partial\Phi(v)$ verifying

$$\langle g - h, x - v\rangle \leqq \gamma \|x - v\|^{q+1}.$$

We observe that weak q-continuity of $\partial\Phi$ does not imply the differentiability of Φ.

By Remark 2.1 the following condition is obviously equivalent to (3.16):

(3.16) For any x in W, $x \neq v$, and any g in $\partial\Phi(x)$, the following holds:

$$|\langle g, y\rangle - \Phi'(v; y)| \leqq \gamma \|x - v\|^q,$$

where y denotes the vector $\|x - v\|^{-1}(x - v)$.

THEOREM 3.5. *Under the assumptions of Theorem* 3.2, *suppose further that there exists a* $q \in]0, 1]$ *such that* $\partial\Phi$ *is weakly q-continuous at* v. *Then the Q-order of convergence of Algorithm* 2.7 *is at least* $q + 1$.

Proof. Without loss of generality, we may assume that the neighborhood W of v for which (3.15) holds is a certain Euclidean ball centered at v. Thus, by Lemma 3.3 and (3.16), for each $x \neq v$ in W, there exist $\hat{x}$ and $\hat{g}$ in $\partial\Phi(\hat{x})$ such that the following conditions hold:

(i) $\hat{x}$ is in W, $\hat{x} \neq v$, and $\|\hat{x} - v\| < \|x - v\|$;

(ii) $|\Phi(x) - \Phi(v) - \langle g, x - v\rangle| \leqq |\langle \hat{g}, y\rangle - \langle g, y\rangle| \, \|x - v\|$, for all g in $\partial\Phi(x)$ and $|\langle \hat{g}, y\rangle - \Phi'(v, y)| \leqq \gamma \|x - v\|^q$, where y represents the vector $\|x - v\|^{-1}(x - v)$.

So, after a few standard calculations we easily obtain the inequality

$$|\Phi(x) - \Phi(v) - \langle g, x - v\rangle| \leqq 2\gamma \|x - v\|^{q+1} \quad \text{for all } g \text{ in } \partial\Phi(x),$$

which, combined with (3.7), completes the proof. □

4. Stability. As was to be expected, "small" perturbations on the unconditional pair (λ, x) as well as on the successive iterates do not interfere with the convergence. This is the content of the next theorem.

THEOREM 4.1. *Let μ be a real, simple eigenvalue of A and v a corresponding eigenvector such that $\Phi(v) = 1$. Then the set of all unconditional pairs with respect to (μ, v) is an open set of $\mathbb{R}^{n+1}$.*

To simplify the proof we introduce the following technical definition and lemma. Also, for simplicity, we will state briefly that a pair is unconditional, meaning that it is unconditional with respect to (μ, v).

We say that the pair (λ, x) has *property-c* if it is an unconditional pair and if there exists a sequence of pairs that are not unconditional and that converge to (λ, x).

LEMMA 4.2. *Suppose that the pair (λ, x) has property-c. Then there exists a sequence $((\lambda_p, x_p))$ in $\mathscr{C}(\lambda, x)$ whose elements are pairs with property-c.*

Proof. We need to show only that if (λ, x) has property-c, then there exists a choice of g in $\partial\Phi(x)$, such that the first iterate (λ', x') obtained by the equation in (N.3) of Algorithm 2.7, calculated with this g, also has property-c.

Let (λ, x) be an unconditional pair. It is obvious that the elements of any sequence in $\mathscr{C}(\lambda, x)$ are also unconditional pairs. On the other hand, by Proposition 2.6 there exists a neighborhood V of (λ, x), where the subdifferential of F is unconditionally invertible. Hence, if the pair has property-c, there exists a sequence (α_k, y_k) in V of nonunconditional pairs, converging to (λ, x) and such that $\partial F(\alpha_k, y_k)$ is unconditionally invertible.

Now, let g_k be a subgradient of Φ at y_k. Denote by (α'_k, y'_k) the first iterate obtained by the algorithm initialized with (α_k, y_k) and with the choice of g_k in step (N.1). Thus

$$\begin{bmatrix} y'_k \\ \alpha'_k \end{bmatrix} := \begin{bmatrix} y_k \\ \alpha_k \end{bmatrix} - J(\alpha_k, y_k, g_k)^{-1} F(\alpha_k, y_k). \tag{4.1}$$

Since (y_k) converges to x, the set $S := \{x, y_1, \cdots, y_k, \cdots\}$, and therefore the set $\partial\Phi(S)$, are compact. Hence, Theorem 24.4 in [6] applies and assures the existence of a subsequence $(g_{k_s})_s$ of (g_k) converging to an element g of $\partial\Phi(x)$. When we take (4.1) into account, it is easily shown that the subsequence $((\alpha'_{k_s}, y'_{k_s}))$ of nonunconditional pairs converges and that its limit is precisely the unconditional pair (λ_1, x_1) given by

$$\begin{bmatrix} x_1 \\ \lambda_1 \end{bmatrix} = \begin{bmatrix} x \\ \lambda \end{bmatrix} - J(\lambda, x, g)^{-1} F(\lambda, x),$$

where $g := \lim_s g_{k_s}$. Thus (λ_1, x_1) has property-c. □

Proof of Theorem 4.1. Let (λ, x) be an unconditional pair. We need to prove only the existence of a neighborhood of (λ, x) whose elements are unconditional pairs.

Seeking a contradiction, we assume that there exists a sequence of nonunconditional pairs converging to (λ, x). If so, (λ, x) has property-c and by the previous lemma (λ, x) will be the limit of a sequence $((\lambda_p, x_p))$ such that, for all p, the pair (λ_p, x_p) has property-c.

On the other hand, any sequence in $\mathscr{C}(\lambda, x)$ converges to (μ, v), since (λ, x) is unconditional. Hence, according to Theorem 3.2, there exists an integer $p_0 > 0$ such that, for $p \geqq p_0$, the pairs (λ_p, x_p) belong to a neighborhood of (μ, v) all of whose elements are unconditional pairs. Thus, for $p \geqq p_0$, a sequence of nonunconditional pairs converging to (x_p, λ_p) cannot exist. This contradicts the fact that (λ_p, x_p) has property-c. □

Finally we remark that if we combine our result (3.13) and the previous theorem, then we have the following complementary result to (3.13).

COROLLARY 4.3. *Let K be any compact set of $\mathbb{R}^n$ such that for any $x \in K$ the following holds: x does not belong to the column space of $\mu I - A$ and $\langle g, v \rangle \neq 0$ for any $g \in \partial\Phi(x)$. Then there is a neighborhood N of μ such that any pair (λ, x) of $N \times K$ is unconditional with respect to (μ, v) or (μ, w) according to whether $\langle g, v \rangle$ is positive or negative.*

Acknowledgment. The author thanks Professor Marques de Sá for helpful discussions on the subject matter of this paper. The results are part of the author's doctoral thesis [7].

REFERENCES

[1] P. M. ANSELONE AND L. B. RALL, *The solution of characteristic value-vector problem by Newton's method*, Numer. Math., 11 (1968), pp. 38–45.

[2] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solutions of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[3] G. PETERS AND J. H. WILKINSON, *Inverse iteration, ill-conditioned equations and Newton's method*, SIAM Rev., 21 (1979), pp. 339–360.

[4] L. B. RALL, *Newton's method for the characteristic value problem $Ax = \lambda Bx$*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 288–293; *Errata*, J. Soc. Indust. Appl. Math., 10 (1962), p. 228.

[5] S. M. ROBINSON AND K. NICKEL, *Computation of the Perron root and vector of a non-negative matrix*, Tech. Summary Report, Mathematics Research Center, University of Wisconsin, Madison, WI, 1970.

[6] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[7] M. C. SANTOS, *Alguns métodos iterativos para o problema de valores próprios*, Ph.D. thesis, Coimbra, Portugal, 1985.

[8] T. YAMAMOTO, *Error bounds for computed eigenvalues and eigenvectors*, Numer. Math., 34 (1980), pp. 189–199.

SIAM J. MATRIX ANAL. APPL.
Vol. 9, No. 4, October 1988

010

LINEAR MATRIX EQUATIONS, CONTROLLABILITY AND OBSERVABILITY, AND THE RANK OF SOLUTIONS*

HARALD K. WIMMER†

Abstract. The equation

$$(*) \qquad \sum_{i,k} f_{ik} A^i X B^k = C$$

is studied. The controllability matrix of (A, C) and the observability matrix of (B, C) yield bounds for the rank of X. If the solution X is unique it can be expressed in the form

$$X = \sum_{i,k} h_{ik} A^i C B^k.$$

The coefficients h_{ik} are determined by an auxiliary equation of type $(*)$, where the right-hand side is a rank one matrix.

Key words. matrix equations, Lyapunov equations

AMS(MOS) subject classification. 15A24

1. Introduction. The starting point for our investigation is the paper by de Souza and Bhattacharyya [3] on the matrix equation

$$AX - XB = C. \tag{1.1}$$

In this note we will study the more general equation

$$\sum_{i=0}^{p-1} \sum_{k=0}^{q-1} f_{ik} A^i X B^k = C \tag{1.2}$$

where A, B, and C are complex matrices of size $p \times p$, $q \times q$, and $p \times q$, respectively, and $f_{ik} \in \mathbb{C}$. Let us first review the results of [3], and at the same time introduce some notation.

The controllability matrix of a pair A and $L \in \mathbb{C}^{p \times t}$ is defined by

$$K(A, L) = (L, AL, \cdots, A^{p-1} L)$$

and the observability matrix of (B, R), $R \in \mathbb{C}^{t \times q}$ is given by

$$D(B, R) = \begin{pmatrix} R \\ RB \\ \vdots \\ RB^{q-1} \end{pmatrix}.$$

The pair (A, L) is called controllable if rank $K(A, L) = p$ and (B, R) is observable if rank $D(B, R) = q$.

THEOREM 1.1 [3]. *Let $C = LR$ be a full-rank factorization and assume that* (1.1) *has a unique solution X; then*

$$\begin{aligned} \operatorname{rank} X &\leqq \min \{\operatorname{rank} K(A, L), \operatorname{rank} D(B, R)\} \\ &= \min \{\operatorname{rank} K(A, C), \operatorname{rank} D(B, R)\}. \end{aligned} \tag{1.3}$$

* Received by the editors June 8, 1987; accepted for publication (in revised form) April 18, 1988.
† Mathematisches Institut, Universität Würzburg, D 8700 Würzburg, Federal Republic of Germany.

In the case where rank $C = 1$ we have equality in (1.3). Part of the next theorem has been proved by Hearon in [5].

THEOREM 1.2 [3]. *Suppose* $C = lr \neq 0$, $l \in \mathbb{C}^{p\times 1}$, $r \in \mathbb{C}^{1\times q}$, *and let* (1.1) *have a unique solution. Then we have*:

(1.4) (a) $\operatorname{rank} X = \min\{\operatorname{rank} K(A, l), \operatorname{rank} D(B, r)\}$.

(b) *In the case* $p = q$ *the solution* X *is nonsingular if and only if* (A, l) *is controllable and* (B, r) *is observable.*

Let $a(z) = a_0 + \cdots + a_{p-1}z^{p-1} + z^p$ be the characteristic polynomial of A, and let $\lambda_1, \cdots, \lambda_p$ be its eigenvalues. Similarly, let $b(z) = b_0 + \cdots + b_{q-1}z^{q-1} + z^q$ and $\mu_1, \cdots, \mu_q$ be the characteristic polynomial and the eigenvalues of B. Then (1.1) has a solution for any C (which is necessarily unique) if and only if

$$\lambda_i - \mu_k \neq 0, \quad i = 1, \cdots, p, \quad k = 1, \cdots, q. \tag{1.5}$$

To $a(z)$ and $b(z)$ we associate the companion matrices

$$F_a = \begin{pmatrix} 0 & \cdots & 0 & -a_0 \\ 1 & \cdots & 0 & -a_1 \\ \cdot & \cdots & \cdot & \cdot \\ 0 & \cdots & 1 & -a_{p-1} \end{pmatrix},$$

$$\hat{F}_b = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \\ -b_0 & -b_1 & \cdots & -b_{q-1} \end{pmatrix} = (F_b)^T.$$

The matrix

$$M_a = \begin{pmatrix} a_1 & a_2 & \cdots & a_{p-1} & 1 \\ a_2 & \cdot & & 1 & \\ \vdots & & \iddots & & \\ a_{p-1} & 1 & & 0 & \\ 1 & & & & \end{pmatrix} \tag{1.6}$$

is a "symmetrizer" of F_a, i.e.,

$$F_a M_a = M_a F_a^T. \tag{1.7}$$

THEOREM 1.3 [3]. (a) *The equation*

$$F_a H - H\hat{F}_b = (1, 0, \cdots, 0)^T(1, 0, \cdots, 0) \tag{1.8}$$

is consistent if and only if (1.5) *holds.*

(b) *The unique solution of* (1.8), *if it exists, is given by*

$$H = -[M_a\, O_{p\times(q-p)}]a(F_b)^{-1} \quad \text{if } q \geqq p \tag{1.9a}$$

or

$$H = b(\hat{F}_a)^{-1}\begin{pmatrix} M_b \\ O_{(p-q)\times q} \end{pmatrix} \quad \text{if } p \geqq q. \tag{1.9b}$$

With the matrix H, any solution of (1.1) can be expressed as a finite sum. As usual the Kronecker product of two matrices $P = (p_{ik})$ and Q is $P \otimes Q = (p_{ik}Q)$.

THEOREM 1.4 [3]. *If* (1.5) *holds and* $C = LR$, *then the unique solution of* (1.1) *is given by*

$$X = \sum_{i=0}^{p-1} \sum_{k=0}^{q-1} h_{ik} A^i C B^k = K(A,L)(H \otimes I) D(B,R) \tag{1.10}$$

where $H = (h_{ik})$ *is the solution of* (1.8).

Can we extend the preceding results to the general equation (1.2)? We will show that Theorems 1.1, 1.3(a), and 1.4 are special cases of more general results. For the equation

$$X - AXB = C \tag{1.11}$$

we will derive an explicit solution that is a counterpart to (1.9) in Theorem 1.3(b). As Jameson's trick [6] does not seem to work for (1.11), our approach has to be different from that of [3]. An analogue of Theorem 1.2 is valid for (1.11), but, as the following example shows, it does not hold in general for (1.2). Consider the equation

$$X - A^2 X B^2 = e_1 e_1^T \tag{1.12}$$

with

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad B = A^T, \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

The unique solution of (1.12) is

$$X = \operatorname{diag}(1,0,1)$$

and

$$\operatorname{rank} X = 2 < \operatorname{rank} K(A,e_1) = \operatorname{rank} D(B,e_1^T) = 3.$$

Hence in this case (1.4) does not hold, although the pairs (A, e_1) and (B^T, e_1) are controllable.

2. The general equation. The following criterion is due to Sylvester [8]. Put

$$f(x,y) = \sum_{i=0}^{p-1} \sum_{k=0}^{q-1} f_{ik} x^i y^k. \tag{2.1}$$

LEMMA 2.1. *The equation*

$$\sum_{i=0}^{p-1} \sum_{k=0}^{q-1} f_{ik} A^i X B^k = C \tag{1.2}$$

has a unique solution for every C *if and only if*

$$f(\lambda,\mu) \neq 0 \tag{2.2}$$

for all eigenvalues λ *of* A *and for all eigenvalues* μ *of* B.

Equations where the matrix C is of rank one are important; we refer the reader to § 3 for two examples.

THEOREM 2.2. *Assume rank* $C = 1$ *and* (A, C) *is controllable and* (B, C) *is observable. Then* (1.2) *is consistent if and only if* (2.2) *holds.*

Proof. Suppose (1.2) has a solution X. We want to show that under the given assumptions (2.2) holds. If u and v are eigenvectors of A and B such that $uA = \lambda u$ and $Bv = \mu v$, then

$$f(\lambda,\mu) uxv = uCv.$$

A full rank factorization of C is of the form $C = lr$ with $l \in \mathbb{C}^{p\times 1}$ and $r \in \mathbb{C}^{1\times q}$, and (A, l) is controllable and (B, r) is observable. Hence (see, e.g., [13]) $u(A - \lambda I) = 0$ implies $ul \neq 0$. Similarly we have $rv \neq 0$. Therefore $uCv \neq 0$, and $f(\lambda, \mu) \neq 0$.

COROLLARY 2.3. *The equation*

$$\sum_{i=0}^{p-1}\sum_{k=0}^{q-1} f_{ik}F_a^i H\hat{F}_b^k = (1,0,\cdots,0)^T(1,0,\cdots,0) \tag{2.3}$$

is consistent if and only if (2.2) *holds.*

The matrix H in (2.3) leads to an explicit representation of solutions.

THEOREM 2.4. *Assume that the solvability condition* (2.2) *is satisfied. Then the unique solution of* (1.2) *is of the form*

$$X = \sum_{i=0}^{p-1}\sum_{k=0}^{q-1} h_{ik}A^iCB^k = K(A,I)(H\otimes C)D(B,I) \tag{2.4}$$

where H is given by (2.3).

Proof. We note that

$$AK(A,I) = K(A,I)(F_a\otimes I),$$

$$D(B,I)B = (\hat{F}_b\otimes I)D(B,I).$$

It is easy to verify that the matrix X in (2.4) is a solution.

Djaferis and Mitter [2] derive the finite series solution (2.4) by an algebraic method, which we describe as follows: Let ψ denote the ideal in $\mathbb{C}[x, y]$ generated by $a(x)$ and $b(y)$, and let $V[g(x, y)]$ be the set of zeros of $g(x, y)$ in $\mathbb{C}^2$. Condition (2.2) can now be expressed in other equivalent forms.

THEOREM 2.5 [2]. *The following statements are equivalent*:

(2.2)
(1) $f(\lambda_\rho, \mu_\sigma) \neq 0, \quad \rho = 1, \cdots, p, \quad \sigma = 1, \cdots, q.$
(2) $V[f(x, y)] \cap V[a(x)] \cap V[b(y)] = \varnothing.$
(3) $f(x, g)$ *is a unit in* $\mathbb{C}[x, y]/\psi.$

In particular, if

$$h(x,y) = \sum_{i=0}^{s}\sum_{k=0}^{m} h_{ik}x^iy^k$$

is a polynomial for which

$$f(x,y)h(x,y) \equiv 1 \pmod{\psi}, \tag{2.5}$$

then

$$X = \sum_{i=0}^{s}\sum_{k=0}^{m} h_{ik}A^iCB^k$$

is the unique solution of (1.2).

THEOREM 2.6. *Under condition* (2.2) *there exists a unique polynomial $h(x, y)$ such that* (2.5) *holds and the degree of h is less than p in x and less than q in y. If H is the solution of* (2.3), *then*

$$h(x,y) = (1,x,\cdots,x^{p-1})H\begin{pmatrix} 1 \\ y \\ \vdots \\ y^{q-1} \end{pmatrix}. \tag{2.6}$$

Proof. The uniqueness of such an h is proved in [2]. We check that the polynomial given by (2.6) satisfies (2.5). We have

$$(1,x,\cdots,x^{p-1})F_a=(1,x,\cdots,x^{p-1})x-a(x)(0,\cdots,0,1).$$

Hence

$$(1,x,\cdots,x^{p-1})F_a^i\equiv(1,x,\cdots,x^{p-1})x^i(\operatorname{mod}\psi).$$

Similarly,

$$\hat{F}_b^k\begin{pmatrix}1\\ y\\ \vdots\\ y^{q-1}\end{pmatrix}\equiv y^k\begin{pmatrix}1\\ y\\ \vdots\\ y^{q-1}\end{pmatrix}(\operatorname{mod}\psi).$$

If we multiply (2.3) from the left by $(1,x,\cdots,x^{p-1})$ and from the right by $(1,y,\cdots,y^{q-1})^T$, we obtain (2.5).

To estimate the rank of X we write C as a product: $C=LR$, $L\in\mathbb{C}^{p\times n}$, $R\in\mathbb{C}^{n\times q}$. Then

$$\begin{aligned}K(A,I)(H\otimes LR)D(B,I)&=K(A,I)(I\otimes L)(H\otimes I)(I\otimes R)D(B,I)\\&=K(A,L)(Y\otimes I)D(B,R).\end{aligned}\tag{2.7}$$

If $C=LR$ is a full rank factorization, then

$$\operatorname{rank}K(A,L)=\operatorname{rank}K(A,C)\quad\text{and}\quad\operatorname{rank}D(B,R)=\operatorname{rank}D(B,C).$$

THEOREM 2.7. *Suppose the equation*

$$\sum_{i=0}^{p-1}\sum_{k=0}^{q-1}f_{ik}A^iXB^k=LR$$

has a unique solution X. Then

$$\operatorname{rank}X\leqq\min\{\operatorname{rank}K(A,L),\operatorname{rank}D(B,R)\}.\tag{2.8}$$

Proof. The bound (2.8) follows immediately from (2.4) and (2.7). We will give a second proof, to be used in § 3. There exist nonsingular matrices S and T such that (see, e.g., [13])

$$S^{-1}AS=\begin{pmatrix}A_1&A_{12}\\0&A_2\end{pmatrix},\quad S^{-1}L=\begin{pmatrix}L_1\\0\end{pmatrix},\quad A_1\in\mathbb{C}^{p_1\times p_1},\quad L_1\in\mathbb{C}^{p_1\times n},$$

$$TBT^{-1}=\begin{pmatrix}B_1&0\\B_{21}&B_2\end{pmatrix},\quad RT^{-1}=(R_1\,0),\quad B_1\in\mathbb{C}^{q_1\times q_1},\quad R_1\in\mathbb{C}^{n\times q_1}$$

where (A_1,L_1) and (B_1^T,R_1^T) are controllable and

$$p_1=\operatorname{rank}K(A,L),\qquad q_1=\operatorname{rank}D(B,R).\tag{2.9}$$

Then

$$S^{-1}CT^{-1}=\begin{pmatrix}L_1R_1&0\\0&0\end{pmatrix}.$$

Let

$$S^{-1}XT^{-1}=\begin{pmatrix}X_1&X_{12}\\X_{21}&X_2\end{pmatrix}$$

be partitioned conformably. The block X_2 is the unique solution of the homogeneous equation

$$\sum_i \sum_k f_{ik} A_2^i X_2 B_2^k = 0.$$

Therefore $X_2 = 0$, which in turn yields $X_{12} = 0$ and $X_{21} = 0$. Hence

$$S^{-1}XT = \begin{pmatrix} X_1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad X_1 \in \mathbb{C}^{p_1 \times q_1},$$

and (2.8) follows from (2.9).

We mention without proof that we can obtain an estimate of type (2.8) for a more general equation

$$\sum_{i=0}^{p-1} A^i X B_i = C. \tag{2.10}$$

For basic facts on (2.10) we refer to [11].

3. $X - AXB = C$ with rank $C = 1$. The equations

$$X - AXB = C \tag{1.11}$$

and $AX - XB = C$ have many features in common. We mention only Roth's removal theorem [9] or parallel results for the Lyapunov matrix equations $A^TX + XA = P$ and $X - A^TXA = P$ in stability and inertia theory (see, e.g., [12]). Therefore we can expect that Theorem 1.2 also holds for (1.11) even if it cannot be extended to the general equation (1.2). The following two examples shall illustrate that the case rank $C = 1$ deserves attention.

Notation. Let

$$N = \begin{pmatrix} 0 & & & & \\ 1 & 0 & \cdot & & \\ & 1 & \cdot & \cdot & \\ & \cdot & \cdot & \cdot & \cdot \\ & & & 1 & 0 \end{pmatrix}_{q \times q} \tag{3.1}$$

be a nilpotent Jordan block. Put

$$P = \begin{pmatrix} 0 & & & 1 \\ & & \iddots & \\ & 1 & & 0 \\ 1 & & & \end{pmatrix}.$$

To the complex polynomial

$$b(z) = b_0 + \cdots + b_{q-1} z^{q-1} + z^q \tag{3.2}$$

we associate

$$\tilde{b}(z) = z^q b(z^{-1}), \quad \text{i.e.,} \tag{3.3}$$

$$\tilde{b}(z) = 1 + b_{q-1} z + \cdots + b_0 z^q. \tag{3.4}$$

Example 1. Let c be a complex polynomial with $\deg c \leqq q$. The Bezoutian $\Gamma = \Gamma(b, c)$ is defined by

$$(1, z, \cdots, z^{q-1})\Gamma\begin{pmatrix} 1 \\ y \\ \vdots \\ y^{q-1} \end{pmatrix} = \frac{b(z)c(y)-b(y)c(z)}{z-y}.$$

For a matrix $X \in \mathbb{C}^{q\times q}$ the following assertions are equivalent [7]:

(i) $X = \Gamma(b, c)$ for some nonzero polynomial c;

(ii) $X - NX\hat{F}_b = (b_1, \cdots, b_{q-1}, 1)^T r$ for some nonzero $r \in \mathbb{C}^{1\times q}$.

Example 2. Let b in (3.2) be a real polynomial. The Schur–Cohn matrix

$$D_b = \Gamma(b, \tilde{b})P,$$

which gives information about the location of the roots of b, satisfies [1]

$$X - F_b X F_b^T = (x_{11}, \cdots, x_{q1})^T (x_{11}, \cdots, x_{q1}).$$

If D_b is nonsingular, then the number of positive (respectively, negative) eigenvalues of D_b is equal to the number of roots of b with modulus less (respectively, greater) than 1.

The main result of this section is a counterpart of (1.9). We need several auxiliary results.

LEMMA 3.1 (see, e.g., [11]). *Assume*

$$\lambda_\rho \mu_\sigma \neq 1, \quad \rho = 1, \cdots, p, \quad \sigma = 1, \cdots, q. \tag{3.5}$$

Then the unique solution of (1.11) *is given by*

$$X = \frac{1}{2\pi i}\oint_\Lambda (zI - A)^{-1} C (I - zB)^{-1}\, dz \tag{3.6}$$

where all the eigenvalues λ_ρ of A are in the interior of the simple closed curve Λ and all the zeros of $\det(I - zB)$ *are outside of Λ.*

LEMMA 3.2. *Let M_a be defined by* (1.6). *Then*

$$(zI - F_a)^{-1}(1, 0, \cdots, 0)^T = \frac{1}{a(z)} M_a (1, z, \cdots, z^{p-1})^T.$$

Proof. We recall (1.7) and note that

$$(1, z, \cdots, z^{p-1})(zI - F_a)^{-1} = (0, \cdots, 0, a(z)).$$

LEMMA 3.3 [4]. *Let Λ be a contour containing all zeros of $a(z)$ in its interior. Then*

$$\frac{1}{2\pi i}\oint_\Lambda \frac{1}{a(z)} M_a (1, z, \cdots, z^{p-1})^T (1, z, \cdots, z^{p-1})\, dz = I. \tag{3.7}$$

We now consider the special case of (2.3).

THEOREM 3.4. *Suppose condition* (3.5) *holds. Assume $p \geqq q$. Then the unique solution H of the equation*

$$H - F_a H \hat{F}_b = (1, 0, \cdots, 0)^T (1, 0, \cdots, 0) \tag{3.8}$$

has the form

$$H = \tilde{b}(F_a)^{-1}\begin{pmatrix} I_Q \\ O_{(p-q)\times q} \end{pmatrix}\tilde{b}(N). \tag{3.9}$$

Proof. With the representation (3.6) we have

$$H=\frac{1}{2\pi i}\oint_{\Lambda}(zI-F_a)^{-1}(1,0,\cdots,0)^T(1,0,\cdots,0)(I-z^{-1}\hat{F}_b)\,dz.$$

From Lemma 3.2 we get

$$H=\frac{1}{2\pi i}\oint_{\Lambda}\frac{1}{a(z)\tilde{b}(z)}M_a(1,\cdots,z^{p-1})^T(z^{q-1},\cdots,1)M_b^T\,dz.$$

Because of (3.5) the polynomials a and $\tilde{b}$ are coprime:

$$ad+\tilde{b}g=1 \tag{3.10}$$

for some $d, g\in\mathbb{C}[z]$. Therefore

$$H=\frac{1}{2\pi i}\oint_{\Lambda}\frac{g(z)}{a(z)}M_a(1,\cdots,z^{p-1})^T(1,\cdots,z^{q-1})PM_b\,dz.$$

Note that

$$\frac{1}{2\pi i}\oint_{\Lambda}\frac{z}{a(z)}(1,\cdots,z^{p-1})^T\,dz=F_a^T\frac{1}{2\pi i}\oint_{\Lambda}\frac{1}{a(z)}(1,\cdots,z^{p-1})^T\,dz.$$

Hence

$$H=g(F_a)\frac{1}{2\pi i}\oint_{\Lambda}\frac{1}{a(z)}(1,\cdots,z^{q-1},\cdots,z^{p-1})^T(1,\cdots,z^{q-1})\,dzPM_b.$$

From (3.10) it follows that $\tilde{b}(F_a)g(F_a)=I$. Lemma 3.3 yields

$$\frac{1}{2\pi i}\oint_{\Lambda}\frac{1}{a(z)}M_a(1,\cdots,z^{q-1},\cdots,z^{p-1})^T(1,\cdots,z^{q-1})\,dz=\begin{pmatrix}I_q\\ O_{(p-q)\times q}\end{pmatrix}.$$

The matrix

$$PM_b\begin{pmatrix}1 & & & \\ b_{q-1} & 1 & 0 & \\ \cdot & \cdots & & \\ b_1 & \cdots & b_{q-1} & 1\end{pmatrix}$$

can be written as

$$PM_b=I+b_{q-1}N+\cdots+b_1N^{q-1}+b_0N^q=\tilde{b}(N).$$

We can also verify directly that the matrix (3.9) is a solution: obviously

$$\tilde{b}(F_a)(1,0,\cdots,0)^T=(1,b_{q-1},\cdots,b_1,0,\cdots,0)^T,$$

$$(1,0,\cdots,0)M_b^{-1}=(0,\cdots,0,1).$$

Equation (3.8) follows from

$$(P,O)^T-F_a(P,O)^TF_b=(1,b_{q-1},\cdots,b_1,0,\cdots,0)^T(0,\cdots,0,1).$$

Hence (3.9) is valid for matrices over an arbitrary field, provided the polynomials a and $\tilde{b}$ are coprime.

With the matrix H in (3.9) we obtain the analogue of Theorem 1.2.

THEOREM 3.5. *Suppose* $C=lr\neq 0$, $l\in\mathbb{C}^{p\times 1}$, $r\in\mathbb{C}^{1\times q}$, *and let*

$$X-AXB=C \tag{1.11}$$

have a unique solution. Then we have the following:

(a) rank $X = \min\{\operatorname{rank} K(A, l), \operatorname{rank} D(B, r)\}$.

(b) *In the case $p = q$ the solution X is nonsingular if and only if (A, l) is controllable and (B, r) is observable.*

Proof. From the second proof of Theorem 2.7, it suffices to prove (a) for the case where (A, l) is controllable and (B, r) is observable. Then the matrices

$$K = (l, Al, \cdots, A^{p-1}l)$$

and

$$D = \begin{pmatrix} r \\ rB \\ \vdots \\ rB^{q-1} \end{pmatrix}$$

are nonsingular,

$$K^{-1}AK = F_a, \quad K^{-1}l = (1, 0, \cdots, 0)^T, \quad DBD^{-1} = \hat{F}_b, \quad rD^{-1} = (1, 0, \cdots, 0).$$

If X is the solution of (1.11), then $H = K^{-1}XD^{-1}$ satisfies (3.8). In the case $p \geqq q$ we have (3.9). Because of $\tilde{b}(0) = 1$ the factor $\tilde{b}(N)$ in (3.9) is nonsingular and rank $H = q$, which completes the proof.

Acknowledgment. This paper was written at the Australian National University in Canberra. I gratefully acknowledge the hospitality of the Department of Mathematics at the Institute of Advanced Studies.

Note added in proof. Theorem 2.4 can already be found in the following paper: N. J. Young, Formulae for the solution of Lyapunov matrix equations, Internat. J. Control, 31 (1980), pp. 159–179.

REFERENCES

[1] B. N. DATTA, *Application of Hankel matrices of Markov parameters to the solutions of the Routh–Hurwitz and the Schur–Cohn problems*, J. Math. Anal. Appl., 68 (1978), pp. 276–290.

[2] T. E. DJAFERIS AND S. K. MITTER, *Algebraic methods for the study of some linear matrix equations*, Linear Algebra Appl., 44 (1982), pp. 125–142.

[3] E. DE SOUZA AND S. P. BHATTACHARYYA, *Controllability, observability and the solution of $AX - XB = C$*, Linear Algebra Appl., 39 (1981), pp. 167–188.

[4] P. A. FUHRMANN, *On symmetric rational transfer functions*, Linear Algebra Appl., 50 (1983), pp. 167–250.

[5] J. Z. HEARON, *Nonsingular solutions of $TA - BC = C$*, Linear Algebra Appl., 16 (1977), pp. 57–63.

[6] A. JAMESON, *Solution of the equation $AX + XB = C$ by inversion of an $M \times M$ or $N \times N$ matrix*, SIAM J. Appl. Math., 16 (1968), pp. 1020–1023.

[7] V. PTÁK, *Lyapunov, Bézout, and Hankel*, Linear Algebra Appl., 58 (1984), pp. 363–390.

[8] J. J. SYLVESTER, *Sur la solution du cas le plus général des équations linéaires en quantités binaires c'est-à-dire en quaternions ou en matrices du second ordre*, C. R. Acad. Sci. Paris, 99 (1884), pp. 117–118.

[9] H. K. WIMMER, *The matrix equation $X - AXB = C$ and an analogue of Roth's theorem*, Linear Algebra Appl., to appear.

[10] H. K. WIMMER AND A. D. ZIEBUR, *Die Lösung der Matrizengleichung $X - AXB = C$ durch Integration*, Elem. Math., 27 (1970), pp. 60–61.

[11] ———, *Blockmatrizen und lineare Matrizengleichungen*, Math. Nachr., 59 (1974), pp. 213–219.

[12] ———, *Remarks on inertia theorems for matrices*, Czechoslovak Math. J., 25 (1975), pp. 556–561.

[13] W. M. WONHAM, *Linear Multivariable Control. A Geometric Approach*, Lecture Notes in Econom. and Math. Systems 101, Springer-Verlag, Berlin, New York, 1974.

SIAM J. MATRIX ANAL APPL
Vol. 9, No. 4, October 1988

FIVE-DIAGONAL TOEPLITZ DETERMINANTS AN THEIR RELATION TO CHEBYSHEV POLYNOMIALS*

ROBERT B. MARR† AND GEORGE H. VINEYARD‡

Abstract. A five-diagonal Toeplitz (5DT) determinant is defined as having zeros everywhere except in its five principal diagonals, with each principal diagonal having the same element in all positions. Thus the determinant depends on five arbitrary parameters in addition to its order. The general 5DT determinant of order n is shown to be given by a simple closed expression involving Chebyshev polynomials of the second kind of order $n+1$. An explicit generating function for the determinants is also derived such that the nth coefficient of a power series expansion of the function is the nth-order five-diagonal Toeplitz determinant.

Key words. determinants, Toeplitz matrices, Chebyshev polynomials

AMS(MOS) subject classifications. primary 15A15; secondary 33A65

1. Introduction. In the course of some work requiring evaluation of multidiagonal determinants of arbitrarily large order, we have found simple, closed expressions for the determinant of an arbitrary five-diagonal Toeplitz (5DT) matrix in terms of Chebyshev polynomials of the second kind. As a byproduct of this result, we have also discovered a generating function for the 5DT determinants, enabling us to extract the nth-order determinant as the nth coefficient in a power series expansion of an explicit function.

The strategy starts with the observation that the product of any two general tridiagonal Toeplitz (3DT) matrices is a five-diagonal (5D) matrix which differs from a Toeplitz one only in its upper left $(1,1)$ and lower right (n,n) elements. The determinant of such an "imperfect" matrix is given by a linear combination of the determinants of three successive orders of "perfect" 5DT determinants, and the determinant of any 3DT matrix can be expressed as a Chebyshev polynomial. There are, moreover, three quite distinct ways of choosing the 3DT matrix factors, so that a set of linear equations can be generated, which are easily solved to yield closed-form expressions for the determinants of interest.

The determination of the possible 3DT matrix factors involves the solution of a cubic equation, the coefficients of which depend only on the five parameters appearing as entries in the determinant and not on the order n. It can be shown that the 5DT determinant of any order is in fact a completely symmetric polynomial in the three roots of this equation, and it is in analyzing the nature of this polynomial dependence that we are led to the generating function which precisely expresses the determinant and its dependence on the original set of five parameters.

2. Definitions and proof of the main result. Let $D_n(a,b,c)$ be the nth-order 3DT determinant,

*Received by the editors March 31, 1986; accepted for publication (in revised form) January 7, 1988. This work was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract DE-AC02-76CH00016.

†Brookhaven National Laboratory, Upton, New York 11973.

‡Dr. Vineyard is now deceased.

$$(1) \qquad D_n(a,b,c) = \begin{vmatrix} a & b & 0 & \cdots & \cdots & 0 \\ c & a & b & 0 & \cdots & 0 \\ 0 & c & a & b & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & c & a & b \\ 0 & \cdots & \cdots & 0 & c & a \end{vmatrix}.$$

By expanding in minors on the first row, we find the recurrence relation,

$$(2) \qquad D_n(a,b,c) = aD_{n-1}(a,b,c) - bcD_{n-2}(a,b,c), \qquad n = 2,3,\cdots.$$

With the initial conditions, $D_1 = a, D_2 = a^2 - bc$, we may employ the recurrence relations for Chebyshev polynomials [2] to obtain, for $n = 1,2,\cdots$,

$$(3) \qquad \begin{aligned} D_n(a,b,c) &= (bc)^{n/2} U_n\left(\frac{a}{2\sqrt{bc}}\right) && \text{if } bc \neq 0 \\ &= a^n && \text{if } bc = 0, \end{aligned}$$

where U_n is the nth-degree Chebyshev polynomial of the second kind, defined by

$$(4) \qquad U_n(Z) = \frac{(Z+\sqrt{Z^2-1})^{n+1} - (Z-\sqrt{Z^2-1})^{n+1}}{2\sqrt{Z^2-1}}.$$

Note that $a^n = \lim_{bc\to 0} (bc)^{n/2} U_n\left(\frac{a}{2\sqrt{bc}}\right)$. Equation (3) can also be proved from Wolstenholme's formula [1]

$$(5) \qquad D_n(a,b,c) = a^n \prod_{j=1}^{n} \left(1 - \frac{2\sqrt{bc}}{a} \cos\frac{j\pi}{n+1}\right).$$

Now define the nth-order "imperfect" 5D determinant,

$$(6) \qquad P_n^{\alpha\beta} = \left(P_n^{\alpha\beta}(x,y,z,v,w)\right) = \begin{vmatrix} x-\alpha & y & v & 0 & \cdots & \cdots & 0 \\ z & x & y & v & 0 & \cdots & 0 \\ w & z & x & y & v & 0 & 0 \\ 0 & w & z & x & y & v & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & w & z & x & y \\ 0 & \cdots & \cdots & 0 & w & z & x-\beta \end{vmatrix}.$$

By appropriate expansion of this determinant we find the relation,

$$(7) \qquad P_n^{\alpha\beta} = P_n - (\alpha+\beta)P_{n-1} + \alpha\beta P_{n-2},$$

where $P_n = P_n^{00}$ is the "perfect" 5DT determinant defined by (6) with $\alpha = \beta = 0$. We note the starting values $P_0 = 1, P_1 = x$, and $P_2^{\alpha\beta} = \begin{vmatrix} x-\alpha & y \\ y & x-\beta \end{vmatrix}$. For later use it is also convenient to define

$$P_1^{\alpha\beta} = x - \alpha - \beta, \quad P_0^{\alpha\beta} = 1, \quad P_{-1} = P_{-2} = 0,$$

so that (7) is valid for all integers $n \geqq 0$.

Next, by matrix multiplication observe that

$$(8) \qquad P_n^{\alpha\beta} = D_n(a,b,c)D_n(a',b',c'),$$

provided the following equations are satisfied:

$$aa' + bc' + cb' = x, \tag{9a}$$

$$ab' + ba' = y, \tag{9b}$$

$$ac' + ca' = z, \tag{9c}$$

$$bb' = v, \tag{9d}$$

$$cc' = w, \tag{9e}$$

$$cb' = \alpha, \tag{9f}$$

$$bc' = \beta. \tag{9g}$$

It is obvious that all of these equations are invariant under the transformation,

$$(a, b, c, a', b', c') \to (\lambda a, \lambda b, \lambda c, \lambda^{-1}a', \lambda^{-1}b', \lambda^{-1}c'), \tag{10}$$

for any $\lambda \neq 0$. Therefore, as expected, those $P_n^{\alpha\beta}$ that are expressible in the form given by (8) involve at most five free parameters, which we take initially as x, y, z, v, and w. (In fact, there are only four free parameters, as will be seen.)

We could treat (9a–e) as a system of equations in a, b, c, a', b', c' yielding classes of solutions equivalent under transformation of the form (10), and then use (9f,g) to compute α and β for each such equivalence class. We prefer here the alternative approach of first eliminating (a, b, c, a', b', c') from the seven equations (9a–g), leaving a pair of equations which can be solved for α and β directly. These equations may then be thought of as "consistency conditions" for (8). One such equation, namely

$$\alpha\beta = vw, \tag{11}$$

follows immediately from (9d–g). To obtain the other, first combine (9b,c) using (9f,g) to get

$$yc - zb = (\alpha - \beta)a, \tag{12a}$$

$$zb' - yc' = (\alpha - \beta)a'. \tag{12b}$$

Next, multiply these two equations and use (9d–g) along with the identity, $(\alpha-\beta)^2 = (\alpha+\beta)^2 - 4\alpha\beta$, and (11) to obtain

$$yzs - y^2w - z^2v = (s^2 - 4vw)aa', \tag{13}$$

where $s = \alpha + \beta$. Finally, substitute into (13) the formula, $aa' = x - s$, obtained from (9a,f,g). The result is a cubic equation in s which, after some rearrangement, has the form,

$$s^3 - xs^2 + (yz - 4vw)s - (y^2w + z^2v - 4xvw) = 0. \tag{14}$$

Straightforward elimination of variables in the set (9) leads to a sixth-degree equation in one variable. The foregoing work shows how this higher-degree equation is equivalent to a cubic equation in properly chosen variables.

In the evaluation of D_n we encounter the combinations $a/\sqrt{bc}$ and $a'/\sqrt{b'c'}$ (see (3)). These can be determined as follows: Multiply (9b) by ab and (9c) by ac, respectively, and employ $bb' = v, cc' = w$, and $aa' = x - s$, to get

$$va^2 + (x - s)b^2 = aby \tag{15}$$

and

$$wa^2 + (x - s)c^2 = acz. \tag{16}$$

Multiplying (15) by (16), dividing by b^2c^2, and employing (9f,g) yields for $A = a^2/(bc)$ the quadratic equation,

$$vwA^2 + [s(x-s) - yz]A + (x-s)^2 = 0. \tag{17}$$

Because of the symmetry between primed and unprimed parameters, $(a')^2/(b'c') = A'$ obeys the same equation. However, A and A' must be distinct roots whenever possible in order that all of the equations (9a–d) will be satisfied, as can easily be shown.

Denote the three roots of (14) by $s_j, j = 1,2,3$. Each s_j gives an equation of the form of (17) for A; each of these will have two roots, which we distinguish by suffixes $+$ and $-$. The corresponding solutions of (17) will thus be denoted $A_j^{\pm}$.

Now, employing (8) and (3) we can write

$$P_{n,j}^{\alpha\beta} = (vw)^{n/2} U_n\left(\sqrt{A_j^+}/2\right) U_n\left(\sqrt{A_j^-}/2\right). \tag{18}$$

Here we have added a subscript to $P_n^{\alpha\beta}$ to indicate (without inconvenience to the printer) which root S_j of (14) has been used — it is then implicit that α and β must satisfy the two conditions, $\alpha\beta = vw$, and $\alpha+\beta = s_j$. Depending on how many distinct roots (14) has, we see that (7) has up to three different versions:

$$P_{n,j}^{\alpha\beta} = P_{n,j} - s_j P_{n-1,j} + vwP_{n-2,j}, \qquad j = 1,2,3. \tag{19}$$

By subtracting any two of these equations, it follows that $(s_k - s_j)P_n = \left(P_{n+1,j}^{\alpha\beta} - P_{n+1,k}^{\alpha\beta}\right)$ for $j,k = 1,2,3$.

With (18), we therefore have the following theorem.

THEOREM 1. *Let $P_n = P_n^{00}$ be the determinant of the five-diagonal Toeplitz matrix with elements x,y,z,v,w, as depicted in* (6). *Assume:* (A) $vw \neq 0$, *and* (B) *the cubic equation* (14) *has at least two distinct roots, s_j and s_k. Then,*

$$P_n = (vw)^{(n+1)/2}(s_k - s_j)^{-1}\left[U_{n+1}\left(\frac{\sqrt{A_j^+}}{2}\right) U_{n+1}\left(\frac{\sqrt{A_j^-}}{2}\right) - U_{n+1}\left(\frac{\sqrt{A_k^+}}{2}\right) U_{n+1}\left(\frac{\sqrt{A_k^-}}{2}\right)\right], \tag{20}$$

where the $A_j^{\pm}$ are the roots of the quadratic equation (17) with $s = s_j$, and where U_{n+1} is the $(n+1)$th-degree Chebyshev polynomial of the second kind.

Remark. Expressions for P_n valid when either or both of the assumptions (A) and (B) are violated can in principle be derived from (20) by taking limits. For the sake of brevity, we omit these derivations. It should be noted, however, that the alternative formulation to be presented in the next section applies with no such restrictive assumptions. The case $vw = 0$ is also presented explicitly in §4.

3. Considerations of symmetry; the generating function. From the form of (17), the three roots s_1, s_2, s_3 must obey the following relations:

$$s_1 + s_2 + s_3 = x, \tag{21a}$$

$$s_1s_2 + s_2s_3 + s_3s_1 = yz - 4vw, \tag{21b}$$

$$s_1s_2s_3 = y^2w + z^2v - 4xvw. \tag{21c}$$

These can be rearranged into the more succinct forms,

$$x = \sum_{j=1}^{3} s_j \tag{22a}$$

and

$$(y\sqrt{w} \pm z\sqrt{v})^2 = \prod_{j=1}^{3} (s_j \pm 2M), \tag{22b}$$

where $M = \sqrt{vw}$.

Now consider the quadratic equation (17) with one of the roots, s_1 (say), substituted for s. Using (21a,b) to eliminate x and yz, we find

$$M^2A^2 - (s_2s_3 + 4M^2)A + (s_2 + s_3)^2 = 0, \tag{23}$$

where (as before) $M^2 = v \cdot w$. The two roots, $A_1^{\pm}$, of this equation therefore depend only on M, s_2 and s_3. Because of (18) we infer that the imperfect determinant $P_{n,j}^{\alpha\beta}$ associated with a particular root s_j of (14), which ostensibly depends on x, y, z, v, and w, can in fact be expressed as a certain function, F_n, of only three variables: M and the *other two roots* of (14).

Using (7), we therefore have, for any $n \geqq 0$,

$$P_n - s_jP_{n-1} + M^2P_{n-2} = F_n(s_k, s_\ell, M), \tag{24}$$

$< s_j, s_k, s_\ell >$ being any permutation of $< s_1, s_2, s_3 >$. Of course, $P_0 = 1$ and $P_1 = x = s_1, +s_2, +s_3$, allowing us to infer that P_n generally depends on only the four free parameters, s_1, s_2, s_3, and M. To make this dependence explicit, define the generating function $\mathcal{P}_\lambda$ by the formal power series

$$\mathcal{P}_\lambda = \sum_{n=0}^{\infty} \lambda^n P_n. \tag{25}$$

Multiplying both sides of (25) by $(1 - s_j\lambda + M^2\lambda^2)$,

$$\begin{aligned}(1 - s_j\lambda + M^2\lambda^2)\mathcal{P}_\lambda &= \sum_{n=0}^{\infty} (1 - s_j\lambda + M^2\lambda^2)\lambda^n P_n, \\ &= \sum_{n=0}^{\infty} \lambda^n (P_n - s_jP_{n-1} + M^2P_{n-2}),\end{aligned}$$

and employing (24), we have

$$\begin{gathered}(1 - s_j\lambda + M^2\lambda^2)\mathcal{P}_\lambda = \sum_{n=0}^{\infty} \lambda^n F_n(s_k, s_\ell; M), \\ j, k, \ell = 1, 2, 3, \qquad j \neq k \neq \ell \neq j.\end{gathered} \tag{26}$$

From the three relations (26) we deduce that $\mathcal{P}_\lambda$ can only have the form

$$\mathcal{P}_\lambda(x, y, z, v, w) = c(\lambda, M) \left[\prod_{j=1}^{3} (1 - s_j\lambda + M^2\lambda^2) \right]^{-1}, \tag{27}$$

where $c(\lambda, M)$ is a function that remains to be determined.

Consider the special case where $v = w = M$, $x = y = z = 0$, and recall the definition $P_0 = 1$. Then we find that $P_{n+4} = M^4 P_n (n \geqq 0)$ and that $P_1 = P_2 = P_3 = 0$. It follows that

$$\begin{aligned} P_n(0,0,0,M,M) &= M^n \quad n \equiv 0 (\text{mod } 4), \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{28}$$

Also, for these particular values of the parameters, (14) reduces to $s^3 - 4M^2 s = 0$, giving

$$s_1 = 0, \quad s_2 = 2M, \quad s_3 = -2M. \tag{29}$$

Equation (25) can now be written, for this special case,

$$\mathcal{P}_\lambda = \sum_{n=0}^{\infty} (\lambda M)^{4n} = \frac{1}{1 - \lambda^4 M^4}. \tag{30}$$

Also, from (29) and (27)

$$\mathcal{P}_\lambda = \frac{c(\lambda, M)}{(1 + M^2\lambda^2)(1 - 2M\lambda + M^2\lambda^2)(1 + 2M\lambda + M^2\lambda^2)}. \tag{31}$$

From (30) and (31) we find $c(\lambda, M) = 1 - M^2\lambda^2$, so with (27) we obtain finally, the following result.

THEOREM 2.

$$\sum_{n=0}^{\infty} P_n \lambda^n = \frac{1 - M^2\lambda^2}{\prod_{j=1}^{3} (1 - s_j\lambda + M^2\lambda^2)}, \tag{32}$$

where s_1, s_2, s_3 are the roots of (14), *and* $M^2 = vw$.

Alternatively, we can employ the relations (21) to write more explicitly,

$$\begin{aligned} \sum_{n=0}^{\infty} P_n \lambda^n = (1 - \lambda^2 w) \\ \cdot \Big[(1 + \lambda^2 vw)^3 - x\lambda(1 + \lambda^2 vw)^2 + (yz - 4vw)\lambda^2(1 + \lambda^2 vw) \\ - \lambda^3(y^2 w + z^2 v - 4xvw) \Big]^{-1} \end{aligned} \tag{33}$$

Remark. The relation (33) can also be obtained from a six-term recursion relation satisfied by the determinant P_n, the derivation of which seems to require a much more lengthy computation. The fact that P_n depends on x, y, z, v, and w only through the four combinations, $x, yz, y^2 w + z^2 v$, and vw, appearing in (33) can be shown more directly by applying elementary row and column operations to the original determinant, (6).

4. Special cases. It is seen from the foregoing that the general five-diagonal Toeplitz determinant can be expressed in terms of Chebyshev polynomials and the roots of two algebraic equations – the first a cubic equation, the second a quadratic. In special situations these results reduce to substantially simpler forms. In the first class of special cases treated, which can be described as "pseudosymmetric," the parameters obey the relation

$$y^2 w = z^2 v. \tag{34}$$

Choose the minus signs in (22b) to find, for this case,

$$\prod_{j=1}^{3} (s_j - 2M) = 0.$$

(Note that choosing the plus sign in (22b) is the same thing as choosing for M the other root of $M^2 = vw$ and does not yield anything essentially different.) Thus, one root of (14) is $2M$. Call this s_1:

$$s_1 = 2M = 2\sqrt{vw}. \tag{35}$$

Then from (21a)

$$s_2 + s_3 = x - 2M \tag{36}$$

and from (21c), after reduction,

$$s_2 s_3 = y^2 \sqrt{\frac{w}{v}} - 2x\sqrt{vw}. \tag{37}$$

Eliminating s_3 between (36) and (37) gives the quadratic equation for s_2,

$$s_2^2 + (2M - x)s_2 + y^2 \sqrt{\frac{w}{v}} - 2x\sqrt{vw} = 0. \tag{38}$$

Notice that the symmetric 5DT is a special example of this, for which $y = z$ and $v = w$. Equations (35) and (38) still hold.

Another subset of the pseudosymmetric case can be referred to as a banded determinant, i.e., the Toeplitz determinant for which

$$y = z = 0.$$

Equations (35) and (38) still apply, with (38) becoming slightly simpler.

Another class of special cases worth mentioning occurs when $M^2 = vw = 0$. Obviously, if both v and w vanish, P_n reduces to the tridiagonal (3DT) case treated at (3). However, if we take $w = 0$ and $v \neq 0$ we have a four-diagonal Toeplitz (4DT) determinant. Correspondingly, we may take $c' = 0$, implying $\beta = 0$ (9e,b), and instead of (18), we now have, from (8) and (3),

$$P_{n,j}^{\alpha 0} = (a_j' m_j)^{n/2} U_n \left(\frac{\sqrt{A_j}}{2} \right), \tag{39}$$

where $m_j = \sqrt{b_j c_j}$, and $A_j = a_j^2/m_j^2$; thus the 4DT determinant can be expressed as a linear combination of Chebyshev polynomials, instead of a bilinear one. In place of (17), A_j now satisfies a linear equation

$$[s_j(x - s_j) - yz]A_j + (x - s_j)^2 = 0, \tag{40}$$

although the roots s_j must still be obtained by solving a general cubic, namely

$$s^3 - xs^2 + yzs - z^2v = 0. \tag{41}$$

The quantity $a'_j m_j$ appearing in (39) can be easily evaluated by returning to (9) and recalling that $m'_j = w = c'_j = \beta = 0$, $s_j = \alpha$. Thus,

$$a'_j m_j = a'_j c_j \sqrt{\frac{bb'}{cb'}} = z\frac{\sqrt{v}}{s_j}. \tag{42}$$

An alternative expression easily derived from (22b) is: $a'_j m_j = \sqrt{s_k s_\ell}$.

REFERENCES

[1] THOMAS MUIR, *The Theory of Determinants*, Vol. IV, Dover, New York, 1960, p. 401.

[2] U.W. HOCHSTRASSER, *Orthogonal Polynomials*, Handbook of Mathematical Functions, M. Abramowitz and I.A. Stegun, eds., Applied Mathematics Series Vol. 55, National Bureau of Standards, Gaithersburg, MD, 1964, Chap. 22, p. 771.

SIAM J. MATRIX ANAL APPL
Vol. 9, No. 4, October 1988

012

A SHARP BOUND FOR PRODUCTS OF HYPERBOLIC PLANE ROTATIONS*

C. T. PAN† AND KERMIT SIGMON‡

Abstract. An algorithm for downdating a least squares problem using hyperbolic plane rotations has recently been presented and analyzed by Alexander, Pan, and Plemmons. Their analysis of the numerical stability of the algorithm rests on the existence of a tight bound on the product of the norms of a certain collection of hyperbolic rotations. The main result of this paper, which was obtained in conjunction with that work, establishes the required tight bound. The inequality established may be of interest in its own right.

Key words. hyperbolic rotations, downdating, least squares, roundoff error

AMS(MOS) subject classifications. primary 65G05; secondary 65F20

1. Introduction.

The objective of this paper, simply stated, is to solve the following problem.

PROBLEM. *Given $a \in \mathbf{R}^n$ with $\|a\|_2 < 1$, find a sharp bound for the product*

$$(1+a_1)\left(1+\frac{a_2}{\sqrt{1-a_1{}^2}}\right)\cdots\left(1+\frac{a_n}{\sqrt{1-a_1{}^2-\cdots-a_{n-1}{}^2}}\right)$$

in terms of $\|a\|_2$ and n.

While this problem may be of interest in other contexts, the results presented here were motivated by and obtained in conjunction with the work of Alexander, Pan, and Plemmons [1] in which an algorithm for downdating a least squares problem using hyperbolic plane rotations,

$$H = \begin{bmatrix} c & -s \\ -s & c \end{bmatrix}, \quad c = \cosh\theta, \quad s = \sinh\theta,$$

is analyzed. The conclusion of [1] on the accuracy of the results of the hyperbolic rotation downdating algorithm, under some simplification, rests on the existence of a tight bound for the product of the norms of a certain collection of hyperbolic plane rotations. As we will show, the question of the existence of such a bound is equivalent to the problem stated above. Our main result establishes such a bound. In addition, certain identities needed in [1] are established.

2. Preliminaries.

In this section, we establish notation as well as some identities which are needed both in the sequel and in [1]. The reader is referred to [1] for the details of how this relates to the least squares downdating problem.

Suppose $R \in \mathbf{R}^{n\times n}$ is upper triangular with positive diagonal and $z, a \in \mathbf{R}^n$ satisfy $a^T R = z^T$ and $\|a\|_2 < 1$. We denote by

$$H_1, H_2, \cdots, H_n \in \mathbf{R}^{(n+1)\times(n+1)}$$

*Received by the editors December 1, 1986; accepted for publication (in revised form) December 21, 1987.

†Department of Mathematical Sciences, Northern Illinois University, DeKalb, IL 60115. The work of this author was supported by U. S. Air Force grant no. AFOSR-83-0225-C, while the author was at North Carolina State University.

‡Department of Mathematics, University of Florida, Gainesville, Florida 32611.

the sequence of hyperbolic plane rotations which, if well-defined, consecutively "zero" the coordinates of z:

$$\begin{bmatrix} R^{(0)} \\ z^{(0)T} \end{bmatrix} := \begin{bmatrix} R \\ z^T \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} R^{(k)} \\ z^{(k)T} \end{bmatrix} := H_k \begin{bmatrix} R^{(k-1)} \\ z^{(k-1)T} \end{bmatrix},$$

where

$$H_k := \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & c_k & & -s_k & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \\ & & & -s_k & & c_k & \end{bmatrix}, \qquad t_k := \frac{z_k{}^{(k-1)}}{r_{kk}},$$

$$c_k := \frac{1}{\sqrt{1 - t_k{}^2}}, \qquad s_k := t_k c_k.$$

We now establish identities which express t_k, c_k, and s_k in terms of a. For the remainder of the paper $\|x\|$ will denote the 2-norm of x defined by $\|x\| := \sqrt{x^T x}$.

LEMMA. *If* $\|a\| < 1$, *then for each* $k = 1, 2, \cdots, n$, *we have the following:*

(1) $z^{(k)} = (z - R^T \alpha_k)/\sqrt{1 - (a_1{}^2 + \cdots + a_k{}^2)}$, *where* $\alpha_k := (a_1, \cdots, a_k, 0, \cdots, 0)^T$.

(2) $t_k = a_k/\sqrt{1 - (a_1{}^2 + \cdots + a_{k-1}{}^2)}$; *hence* $|t_k| \leq \|\alpha_k\| \leq \|a\| < 1$ *so that* H_k *is well-defined.*

(3) $c_k = \sqrt{1 - (a_1{}^2 + \cdots + a_{k-1}{}^2)}/\sqrt{1 - (a_1{}^2 + \cdots + a_k{}^2)}$ *and* $s_k = a_k/\sqrt{1 - (a_1{}^2 + \cdots + a_k{}^2)}$.

(4) $c_1 c_2 \cdots c_n = 1/\sqrt{1 - \|a\|^2}$.

Proof. We first prove (1) by mathematical induction on k. It is convenient to set $\alpha_0 :=$ 0(zero vector) and begin the induction at $k = 0$, where we have that $z^{(0)} = z = (z - R^T \alpha_0)/\sqrt{1 - \|\alpha_0\|^2}$.

Now suppose that $z^{(k-1)} = (z - R^T \alpha_{k-1})/\sqrt{1 - \|\alpha_{k-1}\|^2}$. If we let r_k denote the kth column of R, then since $z_k = r_k{}^T \alpha_k = r_k{}^T \alpha_{k-1} + a_k r_{kk}$, it follows from the induction hypothesis that

$$z_k{}^{(k-1)} = \frac{a_k r_{kk}}{\sqrt{1 - \|\alpha_{k-1}\|^2}}$$

and hence

$$t_k = \frac{z_k{}^{(k-1)}}{r_{kk}} = \frac{a_k}{\sqrt{1 - \|\alpha_{k-1}\|^2}}.$$

Then $|t_k| \leq \|\alpha_k\| \leq \|a\| < 1$ so that c_k and s_k are well-defined; in particular, $c_k = \sqrt{1 - \|\alpha_{k-1}\|^2}/\sqrt{1 - \|\alpha_k\|^2}$ and $s_k = a_k/\sqrt{1 - \|\alpha_k\|^2}$. The induction step is

completed by noting that

$$\begin{aligned} z^{(k)} &= -s_k R_k{}^T + c_k z^{(k-1)} \\ &= \frac{-a_k R_k{}^T + \sqrt{1 - \|\alpha_{k-1}\|^2}\, z^{(k-1)}}{\sqrt{1 - \|\alpha_k\|^2}} \\ &= \frac{-a_k R_k{}^T + (z - R^T \alpha_{k-1})}{\sqrt{1 - \|\alpha_k\|^2}} \\ &= \frac{z - R^T \alpha_k}{\sqrt{1 - \|\alpha_k\|^2}}, \end{aligned}$$

where R_k denotes the kth row of R.

It only remains to observe that the identities for t_k, c_k, and s_k in (2) and (3) were established in the proof of (1) and that (4) follows easily from the identity for c_k. □

3. Source of the problem. In the analysis given in [1] of the numerical stability of the hyperbolic rotation downdating algorithm we are lead to the need for a bound for

$$\left\| \prod_{k=n}^{1} H_k \right\| \leq \prod_{k=1}^{n} \|H_k\|.$$

We would like this bound to be in terms of $\|a\|$ and n since Stewart [3] has shown that the nearness of $\|a\|$ to 1 — in particular, the size of $1/\sqrt{1 - \|a\|^2}$ — is a measure of the condition of the problem.

On the one hand, if $\|a\|$ is near to 1, the above products must be large. To see this note that the $(n+1, n+1)$st entry of $H_n \cdots H_2 H_1$ is $c_1 c_2 \cdots c_n = 1/\sqrt{1 - \|a\|^2}$ and hence $1/\sqrt{1 - \|a\|^2} \leq \|H_n \cdots H_2 H_1\|$. [Thanks go to the referee for pointing this out.]

On the other hand, it is easy to show that

$$\|H_k\| = c_k + |s_k| = c_k(1 + |t_k|),$$

so that

$$\prod_{k=1}^{n} \|H_k\| = (c_1 c_2 \cdots c_n) \prod_{k=1}^{n} (1 + |t_k|) = \frac{1}{\sqrt{1 - \|a\|^2}} \prod_{k=1}^{n} (1 + |t_k|).$$

The products are, therefore, bounded as follows.

THEOREM. *If* $\|a\| < 1$, *then*

$$\frac{1}{\sqrt{1 - \|a\|^2}} \leq \left\| \prod_{k=n}^{1} H_k \right\| \leq \prod_{k=1}^{n} \|H_k\| \leq \frac{1}{\sqrt{1 - \|a\|^2}} \prod_{k=1}^{n} (1 + |t_k|).$$

The factor $1/\sqrt{1 - \|a\|^2}$ is just Stewart's measure of the condition of the problem. It remains, therefore, to find a tight bound for $\prod_{k=1}^{n}(1 + |t_k|)$ in terms of $\|a\|$ and n. We should note that, since $t_k = a_k/\sqrt{1 - (a_1{}^2 + \cdots + a_{k-1}{}^2)}$, this is exactly the problem stated at the beginning of the paper.

Of course, since $|t_k| < 1$ for each k, 2^n is one bound for the product, but not a satisfactory one. We will show below that a much tighter bound exists.

4. The bound.

THEOREM. *If* $\|a\| < 1$, *then*

$$\prod_{k=1}^{n}(1+|t_k|) \leq \left(1+\sqrt{1-\sqrt[n]{1-\|a\|^2}}\right)^n.$$

This bound is sharp with equality being obtained when $|t_1| = |t_2| = \cdots = |t_n|$.

Proof. Define f on $\mathbf{R}^n$ by

$$\begin{aligned} f(a_1, a_2, \cdots, a_n) &:= \prod_{k=1}^{n}(1+|t_k|) \\ &= \prod_{k=1}^{n}\left(1+\frac{|a_k|}{\sqrt{1-({a_1}^2+\cdots+{a_{k-1}}^2)}}\right). \end{aligned}$$

For fixed $\|a\|$, we show that the maximum obtained by f on the n-sphere ${a_1}^2 + {a_2}^2 + \cdots + {a_n}^2 = \|a\|^2$ in $\mathbf{R}^n$ is indeed

$$B_n := \left(1+\sqrt{1-\sqrt[n]{1-\|a\|^2}}\right)^n.$$

Since $f(a_1, a_2, \cdots, a_n)$ is invariant under arbitrary change of signs of $a_1, a_2, \cdots, a_n$, it suffices to consider f on the positive cone of $\mathbf{R}^n$. Therefore, we assume henceforth that $a_i \geq 0$, and hence $t_i \geq 0$, for each i.

We first show that equality is obtained when $t_1 = t_2 = \cdots = t_n$. In this case we have that

$$\begin{aligned} {a_{k+1}}^2 &= {a_1}^2(1-{a_1}^2-\cdots-{a_k}^2) \\ &= {a_1}^2(1-{a_1}^2-\cdots-{a_{k-1}}^2)-{a_1}^2{a_k}^2 \\ &= {a_k}^2-{a_1}^2{a_k}^2 \\ &= {a_k}^2(1-{a_1}^2). \end{aligned}$$

A simple inductive argument gives

$${a_k}^2 = {a_1}^2(1-{a_1}^2)^{k-1}$$

for each k. Hence

$$\begin{aligned} \|a\|^2 &= {a_1}^2+{a_2}^2+\cdots+{a_n}^2 \\ &= {a_1}^2\left[1+(1-{a_1}^2)+(1-{a_1}^2)^2+\cdots+(1-{a_1}^2)^{n-1}\right] \\ &= 1-(1-{a_1}^2)^n \end{aligned}$$

since $0 < 1-{a_1}^2 < 1$. Solving for a_1 yields

$$t_k = t_1 = a_1 = \sqrt{1-\sqrt[n]{1-\|a\|^2}}$$

for each k. Therefore, at this particular point b where $t_1 = \cdots = t_k$, we have that

$$f(b) = \prod_{k=1}^{n}(1+t_k) = \left(1+\sqrt{1-\sqrt[n]{1-\|a\|^2}}\right)^n = B_n.$$

To establish the inequality we proceed by induction on n. For $n = 1$ the result is immediate. We assume the inequality holds for $n - 1$ and show it holds for n. It is convenient to set

$$T_n := \{ x \in \mathbf{R}^n \mid \|x\| = \|a\| \text{ and } x_i \geq 0 \text{ for each i} \},$$
$$T_n{}^\circ := \{ x \in \mathbf{R}^n \mid \|x\| = \|a\| \text{ and } x_i > 0 \text{ for each i} \}.$$

Since f is continuous on the compact set T_n, it must attain a maximum at some point of T_n. We next show that this point must be in $T_n{}^\circ$. If, for any fixed i_0, a_{i_0} is constrained to be zero, the product

$$f(a_1, \cdots, a_n) = \prod_{k=1}^{n} \left(1 + \frac{a_k}{\sqrt{1 - (a_1{}^2 + \cdots + a_{k-1}{}^2)}} \right)$$

reduces to the $(n-1)$st case. It follows from the induction hypothesis that $f(a_1, \cdots, a_n) \leq B_{n-1}$ on the complement of $T_n{}^\circ$ in T_n. But it is shown above that $f(b) = B_n$ at that point b in T_n where $t_1 = \cdots = t_n$. Using derivatives one can easily show that $B_n > B_{n-1}$. We then have that $b \in T_n{}^\circ$, so f must attain its maximum in $T_n{}^\circ$.

We complete the proof by showing that the point in $T_n{}^\circ$ where f attains its maximum must in fact be the point b corresponding to where $t_1 = \cdots = t_n$ at which, as shown above, $f(b) = B_n$.

By the principle of Lagrange multipliers, we have that

$$\nabla f(a_1, \cdots, a_n) = \lambda \nabla g(a_1, \cdots, a_n)$$

for some $\lambda \in \mathbf{R}$ at the point a at which f attains its maximum, where $g(x_1, \cdots, x_n) := x_1{}^2 + \cdots + x_n{}^2$ [2, p.374]. Hence $\frac{\partial f}{\partial a_k} = 2\lambda a_k$ for each k so that

$$\frac{1}{a_1}\frac{\partial f}{\partial a_1} = \frac{1}{a_2}\frac{\partial f}{\partial a_2} = \cdots = \frac{1}{a_n}\frac{\partial f}{\partial a_n} = 2\lambda.$$

It then follows from the following lemma that $t_1 = t_2 = \cdots = t_n$ so the proof is complete. □

LEMMA. *If*

$$\frac{1}{a_k}\frac{\partial f}{\partial a_k} = \frac{1}{a_{k+1}}\frac{\partial f}{\partial a_{k+1}}$$

at a point in $T_n{}^\circ$, *then* $t_k = t_{k+1}$.

Proof. First note that since $t_i = a_i / \sqrt{1 - (a_1{}^2 + \cdots + a_{i-1}{}^2)}$ we have

$$\frac{\partial t_i}{\partial a_k} = \begin{cases} \dfrac{a_k a_i}{(1 - a_1{}^2 - \cdots - a_{i-1}{}^2)^{3/2}} = a_k \dfrac{t_i{}^3}{a_i{}^2} & \text{if } i > k, \\ \dfrac{1}{\sqrt{1 - (a_1{}^2 + \cdots + a_{k-1}{}^2)}} = \dfrac{t_k}{a_k} & \text{if } i = k, \\ 0 & \text{if } i < k. \end{cases}$$

Observe that for fixed k, $(1/a_k)\frac{\partial t_i}{\partial a_k} = t_i{}^3/a_i{}^2$ is the same for all $i > k$ and $(1/a_k)\frac{\partial t_i}{\partial a_k} = 0$ for all $i < k$. Hence corresponding terms in the expansions of $(1/a_k)\frac{\partial f}{\partial a_k}$ and $(1/a_{k+1})\frac{\partial f}{\partial a_{k+1}}$ by the product rule for derivatives agree, except

possibly in the kth and $(k+1)$st terms. Therefore

$$\begin{aligned}\frac{1}{a_{k+1}}\frac{\partial f}{\partial a_{k+1}} - \frac{1}{a_k}\frac{\partial f}{\partial a_k} &= \left[\frac{t_{k+1}}{{a_{k+1}}^2}(1+t_k)(1-{t_{k+1}}^2) - \frac{t_k}{{a_k}^2}(1+t_{k+1})\right]\prod_{\substack{i\neq k\\ i\neq k+1}}(1+t_i)\\ &= \left[\frac{t_{k+1}}{{a_{k+1}}^2}(1+t_k)(1-t_{k+1}) - \frac{t_k}{{a_k}^2}\right]\prod_{i\neq k}(1+t_i).\end{aligned}$$

Since $(1/a_k)\frac{\partial f}{\partial a_k} = (1/a_{k+1})\frac{\partial f}{\partial a_{k+1}}$ we have that

$$\frac{t_k}{{a_k}^2} = \frac{t_{k+1}}{{a_{k+1}}^2}(1+t_k)(1-t_{k+1}).$$

But from the relation ${a_i}^2 = {t_i}^2(1-{a_1}^2-\cdots-{a_{i-1}}^2)$ we easily show that $({t_k}^2/{a_k}^2) = ({t_{k+1}}^2/{a_{k+1}}^2)(1-{t_k}^2)$. It follows that

$$\frac{{t_{k+1}}^2}{{a_{k+1}}^2}(1-{t_k}^2) = \frac{t_k t_{k+1}}{{a_{k+1}}^2}(1+t_k)(1-t_{k+1})$$

so that $t_{k+1}(1-t_k) = t_k(1-t_{k+1})$, and hence $t_{k+1} = t_k$. □

4. Numerical values of the bound. Below we give some numerical values of the bound

$$B_n = \left(1+\sqrt{1-\sqrt[n]{1-\|a\|^2}}\right)^n.$$

It can be seen that B_n does not grow rapidly with n unless $\|a\|$ is very close to one. It is on the basis of this fact and the results of Stewart [3] that the general conclusion is drawn in [1] that the results of the hyperbolic rotation downdating algorithm are accurate unless the problem is ill conditioned.

As noted earlier, 2^n is a bound for the relevant product. Since $B_n = 2^n$ when $\|a\| = 1$, the rightmost column of Table 1 below gives approximations of 2^n. It is seen, therefore, that B_n is a significantly tighter bound than 2^n unless $\|a\|$ is very near 1.

TABLE 1
Approximate values of B_n

$\|a\|$	$1-10^{-1}$	$1-10^{-2}$	$1-10^{-4}$	$1-10^{-6}$	$1-10^{-8}$	1
n						
10	27	91	280	482	649	1024
20	144	1138	1×10^4	4×10^4	9×10^4	1×10^6
30	524	8103	2×10^5	1×10^6	5×10^6	1×10^9
40	1553	4266	2×10^6	2×10^7	1×10^8	1×10^{12}
60	1×10^4	7×10^4	1×10^8	4×10^9	5×10^{10}	1×10^{18}
80	5×10^4	7×10^6	4×10^9	3×10^{11}	6×10^{12}	1×10^{24}
100	2×10^5	6×10^7	8×10^{10}	1×10^{13}	5×10^{14}	1×10^{30}

REFERENCES

[1] S. T. ALEXANDER, C. T. PAN, and R. J. PLEMMONS, *Analysis of a recursive least squares hyperbolic rotation algorithm for signal processing*, Linear Algebra Appl. 98 (1988), pp. 3–40.
[2] W. FULKS, *Advanced Calculus*, Third edition, John Wiley, New York.
[3] G. W. STEWART, *Effects of rounding error on an algorithm for downdating the Cholesky factorization*, J. Inst. Math. Appl. 23 (1979), pp. 203–212.